

FORSCHUNGSBERICHTE

aus dem

**Institut für Statistik und Wissenschaftstheorie
der Universität München**

SCHÄTZPROBLEME BEI
SYMMETRISCHEN STICHPROBEN

Schätzung des Erwartungswertes von
Summen von Zufallsvariablen mit
Anwendung auf Inferenz aus Graphen

von
Peter Kremser
München
—1982

1983

Ludwig-Maximilians-Universität München
Institut für Statistik und Wissenschafts-
theorie, Ludwigstraße 33, 8000 München 22



Digitized by the Internet Archive
in 2026

https://archive.org/details/IA41555209_0029

SCHÄTZPROBLEME BEI SYMMETRISCHEN STICHPROBEN

Schätzung des Erwartungswertes von Summen von
Zufallsvariablen mit Anwendung auf Inferenz aus
Graphen

Inaugural-Dissertation
zur Erlangung des Grades Doctor rerum politicarum
(Dr.rer.pol.)
an der Ludwig-Maximilians-Universität München

vorgelegt von
Peter Kremser

1982



Referent:	Professor Dr. A.-R. Wulsten
Korreferent:	Professor Dr. R. Ziegler
Promotionsabschluß- beratung:	21. 2. 1983

VORWORT

Diese Arbeit ist während meiner Tätigkeit am Institut für Statistik und Wissenschaftstheorie entstanden.

Zu besonderem Dank bin ich Herrn Professor Dr. A.-R. Wulsten für seine stets wohlwollende Unterstützung und wertvollen Ratschläge verpflichtet. Gesprächen mit Herrn Professor Dr. H. Stenger und seinen Arbeiten verdanke ich wesentliche Hinweise zum theoretischen Teil der Arbeit. Herrn Professor Dr. Ziegler danke ich für wichtige Anregungen und Hinweise zum Anwendungsbezug. Bei meinem Kollegen Dr. K. Haagen möchte ich mich für wertvolle Diskussionen bedanken. Nicht zuletzt gilt mein Dank Frau G. Clevenz für das Schreiben des Manuskriptes.

München, im Juli 1982

INHALTSVERZEICHNIS

	Seite
<u>Einleitung und Übersicht</u>	1
I. <u>Schätzprobleme bei Zufallsvariablen, die einzelnen Untersuchungseinheiten zugeordnet sind</u>	
1.1 Symmetrische Stichprobenverfahren	7
1.2 Grundbegriffe der Schätztheorie	21
1.3 Die Schätzung des Erwartungswertes der Summe der Zufallsvariablen	30
1.4 Zulässigkeit, vollständige Klassen von Schätzfunktionen	41
1.5 Die Schätzung anderer Parameter	58
II. <u>Schätzprobleme bei Zufallsvariablen, die Paaren von Untersuchungseinheiten zugeord- net sind</u>	
2.1 Problemstellung, Voraussetzungen, Be- zeichnungen	65
2.2 Die Schätzung des Erwartungswertes der Summe der Zufallsvariablen	70
2.3 Zulässigkeit, vollständige Klassen von Schätzfunktionen	85
2.4 Die Schätzung anderer Parameter	100
2.5 Schätzprobleme bei Zufallsvariablen, die sowohl einzelnen als auch Paaren von Untersuchungseinheiten zugeordnet sind	106

III. Anwendung der Ergebnisse auf Inferenz aus
Graphen mit besonderer Berücksichtigung
sozialwissenschaftlicher Fragestellungen

3.1	Graphen als Modell in den Sozial- wissenschaften	117
3.2	Zufallsgraphen	137
3.3	Stichproben aus Graphen, Problem- stellung	145
3.4	Anwendung der Ergebnisse aus Kapitel I	160
3.5	Anwendung der Ergebnisse aus Kapitel II, der Stichprobengraph SI	166
3.6	Anwendung der Ergebnisse aus Kapitel II, der Stichprobengraph SVN	176
3.7	Vergleich einiger Stichprobengraphen	182
3.8	Bemerkungen zum "network sampling" von Granovetter	191

<u>Literaturverzeichnis</u>	204
-----------------------------	-----

<u>Anhang</u>	Vergleich von Bezeichnungen aus Kap. I, II und III	211
---------------	---	-----

EINLEITUNG UND ÜBERSICHT

Die Stichprobentheorie stellt Methoden bereit, um aus der Beobachtung eines Teils einer endlichen Population von Untersuchungseinheiten auf Eigenschaften der ganzen Population zu schließen. In einem weiteren Sinne trifft diese Kennzeichnung auf die gesamte schließende Statistik zu, wobei allerdings die Stichprobentheorie "reale" Populationen beschreiben will, während die meisten Gebiete der Statistik erst bei hypothetischen, durch Wahrscheinlichkeitsverteilungen charakterisierten Gesamtheiten ansetzen. Den Untersuchungseinheiten sind Werte zugeordnet, die überwiegend als Konstanten aufgefaßt werden. Durch einen Zufallsmechanismus wird eine Stichprobe aus der Menge aller Untersuchungseinheiten ausgewählt und aus den beobachteten Werten der Stichprobe werden Kennzahlen der Häufigkeitsverteilung in der Grundgesamtheit geschätzt. Deshalb basiert die Inferenz auf der Stichprobenverteilung, die durch den Zufallsmechanismus der Auswahl erzeugt wird.

Der Ansatz dieser Arbeit geht in 2 Punkten über die gängige Stichprobentheorie hinaus: die Werte werden nicht als Konstanten, sondern als Realisierungen von Zufallsvariablen aufgefaßt und darüber hinaus können nicht nur den Untersuchungseinheiten, sondern auch Paaren von Untersuchungseinheiten Werte bzw. Zufallsvariablen zugeordnet sein. Behandelt wird im wesentlichen die Schätzung des Erwartungswertes der Summe dieser Zufallsvariablen, wobei in den ersten beiden theoretischen Kapiteln, angeregt durch einen Aufsatz von Stenger (1979), das Schwergewicht auf der Untersuchung der Zulässigkeit von linearen, symmetrischen Schätzfunktionen liegt. Im dritten Kapitel werden die Bedeutung des obengenannten Modells bei der Analyse sozialer Netzwerke dargelegt und Anwendungsmöglichkeiten der vorher

hergeleiteten Ergebnisse diskutiert. Ausgangspunkte dieses Kapitels sind hauptsächlich die Arbeiten von Frank (1971) (1977a) zur Inferenz aus Graphen und die an der sozialwissenschaftlichen Anwendung orientierte Arbeit von Granovetter (1976) zum "network sampling".

Vor einer ausführlicheren Übersicht über die vorliegende Arbeit ist eine Erläuterung der Auffassung der Werte als Realisierungen von Zufallsvariablen notwendig. In der Stichprobentheorie wird dieses Modell als Superpopulation bezeichnet (vgl. Biyani (1978), Cochran (1965), Royall (1970)), wobei hinter dieser Begriffsbildung die Vorstellung steht, daß eine endliche Grundgesamtheit durch eine Zufallsstichprobe aus einer unendlichen "Superpopulation" ausgewählt wird. Den Elementen der endlichen Grundgesamtheit sind also Zufallsvariablen zugeordnet und die Superpopulation wird durch die gemeinsame Wahrscheinlichkeitsverteilung dieser Zufallsvariablen beschrieben. Parameter dieser Verteilung sollen geschätzt werden; z.B. der Erwartungswert der Summe der Zufallsvariablen und nicht die Summe einer Realisierung der Zufallsvariablen. Der in der Stichprobentheorie vorherrschende Fall von konstanten Variablenwerten ist Spezialfall dieses Modells, wenn nur Wahrscheinlichkeitsverteilungen betrachtet werden, die jeweils einer Ausprägung der Zufallsvariablen das Maß 1 geben. Die Motivation dieser Verallgemeinerung wird in dieser Arbeit in Kapitel III, insbesondere in 3.2 und 3.3.3 dargelegt; hier wollen wir uns auf ein Zitat von Royall (1970, S. 386) beschränken:

"Many workers in finite population sampling theory seem to have had a strong inclination to treat the fixed constants $y_1, \dots, y_N$ of interest in any given finite population as fixed constants. ... However, it seems frequently to be true that at some time before $y_1, \dots, y_N$ are fixed it is natural and generally acceptable to consider these

numbers as values, to be realized, of random variables $Y_1, \dots, Y_N$. What particular values $y_1, \dots, y_N$ will appear is uncertain, and this uncertainty can be described probabilistically. Although subjectivists would presumably accept these statements (perhaps they would prefer to eliminate the word frequently), it seems that in many finite populations such probability models are precisely as objective as those generally used by frequentists elsewhere in statistics."

Das erste Kapitel dieser Arbeit führt allgemeine Begriffe ein und behandelt Schätzprobleme für Parameter der Verteilung der Zufallsvariablen, die einzelnen Untersuchungseinheiten zugeordnet sind. In dieser Arbeit gehen wir von symmetrischen Stichproben aus, die in Abschnitt 1.1 charakterisiert werden. Eine Einführung in Grundbegriffe der Schätztheorie wird in Abschnitt 1.2 aus folgenden Gründen durchgeführt: Lehrbücher der Schätztheorie berücksichtigen meistens nicht explizit das Problem der Auswahl von Elementen aus einer endlichen Grundgesamtheit und Monographien der Stichprobentheorie gebrauchen selten eine an der Entscheidungstheorie orientierte Begriffsbildung.

Der Kern des Kapitels besteht in den Abschnitten 1.3 und 1.4: für lineare, symmetrische Schätzfunktionen wird allgemein unter Zugrundelegung der quadratischen Verlustfunktion die Risikofunktion berechnet und unter dem allgemeinen Modell der Superpopulation ohne Einschränkungen an den Wertebereich der Zufallsvariablen die Menge der zulässigen Schätzfunktionen bestimmt. Bei Restriktionen an den Wertebereich kann nur eine Teilmenge der zulässigen Funktionen und eine allerdings nicht wesentlich größere vollständige Klasse von Schätzfunktionen angegeben werden und daher ist Stengers Ergebnis (1979) nicht direkt aus den hier gegebenen Aussagen abzuleiten. Stenger berechnet die Klasse

der zulässigen Schätzfunktionen für positive, feste Variablenwerte in der Grundgesamtheit.

Der Statistiker ist meistens daran interessiert, auch andere Parameter als den Erwartungswert der Summe bzw. des arithmetischen Mittels der Werte der Grundgesamtheit zu schätzen. Der Abschnitt 1.5 zeigt Möglichkeiten zur erwartungstreuen Schätzung anderer Parameter auf, z.B. der Streuung der Werte, und diskutiert Probleme bei der Schätzung der Varianz bzw. des Risikos von Schätzfunktionen unter dem Modell der Superpopulation.

In den ersten 4 Abschnitten des zweiten Kapitels werden die entsprechenden Untersuchungen für Zufallsvariablen durchgeführt, die Paaren von Untersuchungseinheiten zugeordnet sind. Werte von Paaren werden hierbei beobachtet, falls beide Elemente in der Stichprobe sind. Bei der Angabe der Risikofunktion und einer vollständigen Klasse von Schätzfunktionen benötigt man gegenüber Kapitel I einen Parameter mehr. Das Ergebnis zur Zulässigkeit von Schätzfunktionen ist schwächer: es kann nur eine Teilmenge der zulässigen Schätzfunktionen und eine relativ kleine vollständige Menge von Schätzfunktionen angegeben werden und man erhält deshalb nur die Aussage, daß die Menge der zulässigen Schätzfunktionen zwischen den angegebenen Mengen liegen muß.

Die Ausführungen zur Zulässigkeit von Schätzfunktionen sind vor allem für den theoretischen Statistiker von Bedeutung. Für Anwender wird das Hauptinteresse in einem Teilergebnis liegen: es gibt gleichmäßig bessere, allerdings nicht erwartungstreue Schätzfunktionen als die von Frank benutzten Horvitz-Thompson-Schätzfunktionen, und in der Klasse der linearen, symmetrischen Schätzfunktionen sind die einzigen zulässigen erwartungstreuen Schätzfunktionen einfach und intuitiv naheliegend. Bei der Schätzung des Erwartungswertes des

arithmetischen Mittels der Zufallsvariablen wählt man beispielsweise das arithmetische Mittel der beobachteten Werte. Aus diesem Grunde betrachten wir in der Erörterung der Anwendungsmöglichkeiten ausschließlich diese Schätzfunktionen.

Bisher wurden die Summen der Variablen, die Untersuchungseinheiten und Paaren von Untersuchungseinheiten zugeordnet sind, getrennt behandelt. Zur simultanen Schätzung der Erwartungswerte beider Summen wird in Abschnitt 2.5 die Kovarianz der Schätzfunktionen angegeben und die Schätzung von Linearkombinationen beider Parameter untersucht. Das letztgenannte Problem spielt bei der Inferenz aus Graphen eine bedeutende Rolle und außerdem ist die Schätzung der Streuung ein häufig betrachtetes Beispiel dafür.

Die Beschreibung der Anwendungsmöglichkeiten in Kapitel III beschränkt sich auf Inferenz aus Graphen mit besonderer Orientierung an Problemen der Sozialwissenschaften. Dem Verfasser erschien eine Änderung der Darstellungsweise sinnvoll: während die Abschnitte in den ersten beiden theoretischen Kapiteln durch Sätze, Bemerkungen usw. strukturiert sind, werden sie im letzten Kapitel in Teilabschnitte untergliedert. Um diese Arbeit in das bedeutende Gebiet der Analyse sozialer Netzwerke einzuordnen, wird in den Abschnitten 3.1 - 3.3 die Einführung der notwendigen Grundbegriffe der Graphentheorie mit ausführlichen Hinweisen auf sozialwissenschaftliche Inhalte und auf wichtige Literatur verbunden. Der Abschnitt 3.3 stellt hierbei die Überleitung zu dem behandelten Problemkreis dar: den Stichproben aus Graphen.

Bei der Erörterung der Stichprobengraphen wird hauptsächlich gezeigt, wie weit die von Frank gestellten Probleme mit den Aussagen der ersten beiden Kapitel behandelt werden können. Auf spezielle Verteilungsannahmen wird nur insofern

eingegangen, als die Besonderheiten der Parameter bei unbewerteten Graphen beschrieben werden. Es sei allerdings darauf hingewiesen, daß weitere Verteilungsannahmen in den ersten beiden Kapiteln diskutiert werden. Da in den drei einleitenden Abschnitten des Kapitels III auf die sozialwissenschaftliche Bedeutung der Parameter in größerem Maße eingegangen wurde, wird die Anwendung der Ergebnisse auf die Schätzung spezieller Kennzahlen sozialer Netzwerke nicht mehr erläutert.

Die Motivation für die beiden letzten Abschnitte 3.7 und 3.8 ist ein empirisches Problem der Soziometrie, das Grundlage von Granovetters Arbeit (1976) zum "network sampling" ist. Bei Umfragen, aus denen die soziale Dichte oder die durchschnittliche Anzahl der "Bekannten" pro Person ermittelt werden soll, hat man, ohne auf speziellere Operationalisierungen einzugehen, zwei Möglichkeiten: man fragt einfach nach der Anzahl oder legt den Befragten eine Liste vor, auf der sie ihre Bekannten ankreuzen sollen. Falls man beides für gleichwertige Messungen desselben empirischen Befundes hält, führt das Problem der Wahl einer Befragungsart zum Vergleich von Stichprobengraphen. Die realistischere Vorstellung, daß zumindest bei der Frage nach der Anzahl viele Bekannte vergessen werden, wird in Abschnitt 3.8 anhand eines stochastischen Fehlermodells diskutiert.

Der Anhang bietet eine Hilfe zum Vergleich der Bezeichnungen aus Kap. I und II mit den Bezeichnungen der Graphentheorie in Kap. III.

I. SCHÄTZPROBLEME BEI ZUFALLSVARIABLEN, DIE
EINZELNEN UNTERSUCHUNGSEINHEITEN ZUGEORDNET
SIND.

1.1 Symmetrische Stichprobenverfahren

In dieser Arbeit wird von Stichprobenverfahren ausgegangen, die bei der Auswahl nicht bestimmte Untersuchungseinheiten bevorzugen und die gängigen Auswahlmethoden "Ziehen mit und ohne Zurücklegen" umfassen: den symmetrischen Stichprobenverfahren. Sie werden in der neueren Literatur z.B. von Stenger (1979), Stenger und Gabler (1981) und von Sinha (1976) (unter der Bezeichnung "symmetrisch invariant") diskutiert.

Der Kern dieses Abschnittes besteht in der Kennzeichnung der symmetrischen Stichprobenverfahren als Mischung von Wahrscheinlichkeitsverteilungen, die durch das "Ziehen ohne Zurücklegen" erzeugt werden (Satz 1.1.8). Diese Charakterisierung wird in der vorliegenden sowie in der Arbeit von Stenger (1979) bei den Untersuchungen der Zulässigkeit gebraucht. Weiterhin werden eine andere Möglichkeit zur expliziten Angabe von symmetrischen Stichprobenverfahren (Satz 1.1.2) angegeben und später wichtige Formeln zur Verteilung der Stichprobenindikatorvariablen berechnet.

Definition 1.1.1

- (i) Zu einer endlichen Grundgesamtheit Ω mit $|\Omega|=n$ heißt jede Teilmenge $S \subset \Omega$ Stichprobe. Eine Wahrscheinlichkeitsverteilung P auf der Potenzmenge von Ω heißt (Zufalls-) Stichprobenverfahren.

- (ii) Ein Stichprobenverfahren heißt symmetrisch, wenn für gleich mächtige Stichproben S und $S' \subset \Omega$ (d.h. $|S| = |S'|$) gilt:

$$(1.1.1) \quad P(\{S\}) = P(\{S'\})$$

Die Elementarereignisse dieses Wahrscheinlichkeitsraums sind Teilmengen $S \subset \Omega$ und für $P(\{S\})$ werden wir auch vereinfachend $P(S)$ schreiben, wobei beachtet werden muß, daß dann in scheinbarem Gegensatz zu den Axiomen der Wahrscheinlichkeitsrechnung $P(\emptyset) = P(\{\emptyset\}) > 0$ gelten kann.

Mit der Definition der Stichprobe als Teilmenge sind geordnete Stichproben von den Betrachtungen ausgeschlossen. Bei Ziehungsverfahren mit Zurücklegen wird manchmal auch die Anzahl berücksichtigt, mit der eine bestimmte Untersuchungseinheit $i \in \Omega$ gezogen wird; sie geht ebenfalls in den Teilmengenbegriff nicht ein. Der Ausschluß der beiden obengenannten Möglichkeiten ist durch ein Ergebnis von Roy und Chakravarti (1960, S. 395) gerechtfertigt, das beweist, daß Informationen über die Reihenfolge der ausgewählten Elemente oder über die Häufigkeit, mit der sie gezogen wurden, nicht zu besseren Schätzfunktionen führen.

Wir wollen darauf hinweisen, daß wir die Wörter "Stichprobenverfahren" und "Auswahlverfahren" synonym gebrauchen, auch wenn in der Literatur (z.B. Stenger (1972a)) der Begriff "Stichprobenverfahren" anders als hier erst im Zusammenhang mit Schätzfunktionen definiert wird. Eine Stichprobe, die durch ein symmetrisches Auswahlverfahren erzeugt wird, nennen wir kurz symmetrische Stichprobe.

Als Grundgesamtheit wählen wir stets $\Omega = \{1, 2, \dots, n\}$ und definieren zu $i \in \Omega$ die Stichprobenindikatorvariablen:

$$(1.1.2) \quad S_i := \begin{cases} 1 & \text{falls } i \in S \\ 0 & \text{sonst} \end{cases}$$

Der Stichprobenumfang $K := \sum_{i=1}^n S_i$ ist im allgemeinen eine Zufallsvariable und mit kleinem Buchstaben k wird die Realisierung von K , also der aktuelle Stichprobenumfang bezeichnet.

Im folgenden Satz sollen symmetrische Stichprobenverfahren durch Wahrscheinlichkeiten p_m , daß bestimmte Mengen $\Omega_m = \{i_1, \dots, i_m\}$ in der Stichprobe S enthalten sind, charakterisiert werden.

Satz 1.1.2

- (i) Für ein symmetrisches Stichprobenverfahren und für vorgegebene Mengen $\Omega_m \subset \Omega$ mit $|\Omega_m| = m$ gilt:

$$(1.1.3) \quad P(\Omega_m \subset S) = \sum_{k=m}^n \binom{n-m}{k-m} P(S^{(k)}) =: p_m,$$

wobei $S^{(k)}$ beliebig gewählte k -elementige Obermengen von Ω_m sind (d.h. $|S^{(k)}| = k$ und $\Omega_m \subset S^{(k)}$).

- (ii) Aus der Gültigkeit von $P(\Omega_m \subset S) = p_m$ für $m=1, \dots, n$ und alle Ω_m mit $|\Omega_m| = m$ folgt für verschiedene Einheiten $i_1, \dots, i_m, i_{m+1}, \dots, i_{m+\ell}$ mit $1 \leq m+\ell \leq n$:

$$(1.1.4) \quad P(S_{i_1} = \dots = S_{i_m} = 1 \text{ und } S_{i_{m+1}} = \dots = S_{i_{m+\ell}} = 0) \\ = \sum_{j=0}^{\ell} (-1)^j \binom{\ell}{j} p_{j+m}, \text{ wobei } p_0 := 1.$$

(iii) Die Größen $p_m = P(\Omega_m cS)$, $m=1, \dots, n$ bestimmen das symmetrische Stichprobenverfahren, wobei $S \subset \Omega$ mit $|S|=k$ gilt.

$$(1.1.5) \quad P(S) = \sum_{m=n}^n (-1)^{m-k} \binom{n-k}{m-k} p_m$$

Beweis

(von (i)) Da alle $S^{(k)} \subset \Omega$ mit $|S^{(k)}|=k$ die gleiche Wahrscheinlichkeit haben und es $\binom{n-m}{k-m}$ k -elementige Obermengen von Ω_m gibt, gilt:

$$\begin{aligned} (1.1.6) \quad P(\Omega_m cS) &= P\left(\bigcup_{k=1}^n \{S: |S|=k \text{ und } \Omega_m cS\}\right) \\ &= \sum_{k=m}^n P\left(\{S: |S|=k \text{ und } \Omega_m cS\}\right) \\ &= \sum_{k=m}^n \binom{n-m}{k-m} P(S^{(k)}) \end{aligned}$$

Der Wert von $P(S^{(k)})$ hängt laut Voraussetzung nur von k ab und daher hängt der Wert von $P(\Omega_m cS)$ von der Wahrscheinlichkeitsverteilung P nur über die Größe von m ab.

(von (ii)) Für $m=0$ und $\ell \geq 1$ gilt:

$$\begin{aligned} (1.1.7) \quad P(S_{i_1} = \dots = S_{i_\ell} = 0) &= 1 - P(S_{i_1} = 1 \text{ oder } \dots \text{ oder } S_{i_\ell} = 1) \\ &= 1 - \sum_{j=1}^{\ell} (-1)^{j-1} \binom{\ell}{j} P(\{i_1, \dots, i_j\} cS) \\ &\quad \text{(vgl.: Feller (1962) Bd. 1, S. 89)} \\ &= \sum_{j=0}^{\ell} (-1)^j \binom{\ell}{j} p_j. \end{aligned}$$

Für $m \geq 1$ wird die Aussage mittels vollständiger Induktion über ℓ bewiesen. Aus $\ell=0$ folgt:

$$(1.1.8) \quad P(S_{i_1} = \dots = S_{i_m} = 1) = p_m = \sum_{j=0}^0 (-1)^j \binom{0}{j} p_{j+m}.$$

Aus der Gültigkeit von Gl. (1.1.4) für ℓ folgt für $\ell+1$:

$$(1.1.9) \quad P(S_{i_1} = \dots = S_{i_m} = 1 \text{ und } S_{i_{m+1}} = \dots = S_{i_{m+\ell+1}} = 0)$$

$$= P(S_{i_1} = \dots = S_{i_m} = 1 \text{ und } S_{i_{m+1}} = \dots = S_{i_{m+\ell}} = 0)$$

$$- P(S_{i_1} = \dots = S_{i_m} = S_{i_{m+\ell+1}} = 1$$

$$\text{und } S_{i_{m+1}} = \dots = S_{i_{m+\ell}} = 0)$$

$$= \sum_{j=0}^{\ell} (-1)^j \binom{\ell}{j} p_{j+m} - \sum_{j=0}^{\ell} (-1)^j \binom{\ell}{j} p_{j+m+1}$$

(laut Induktionsvoraussetzung)

$$= p_m + \sum_{j=1}^{\ell} (-1)^j \binom{\ell}{j} p_{j+m} + \sum_{j=1}^{\ell} (-1)^j \binom{\ell}{j-1} p_{j+m}$$

$$+ (-1)^{\ell+1} p_{m+1+\ell}$$

$$= (-1)^0 p_m + \sum_{j=1}^{\ell} (-1)^j \left[\binom{\ell}{j} + \binom{\ell}{j-1} \right] p_{j+m}$$

$$+ (-1)^{\ell+1} p_{m+\ell+1}$$

$$= \sum_{j=0}^{\ell+1} (-1)^j \binom{\ell+1}{j} p_{j+m}.$$

(von (iii)) Aus Teil (ii) von Satz 1.1.2 folgt für jede k -elementige Teilmenge $S = \{i_1, \dots, i_k\}$ von Ω :

$$(1.1.10) \quad P(S) = P(S_{i_1} = \dots = S_{i_k} = 1 \text{ und } S_{i_{k+1}} = \dots = S_{i_n} = 0)$$

$$= \sum_{j=0}^{n-k} (-1)^j \binom{n-k}{j} p_{k+j}$$

$$= \sum_{m=k}^n (-1)^{m-k} \binom{n-k}{m-k} p_m$$

Bemerkung 1.1.3

Aus Gl. (1.1.5) läßt sich die Verteilung des Stichprobenumfangs K bestimmen:

$$(1.1.11) \quad P\left(\sum_{i=1}^n S_i = k\right) = \frac{\sum_{\{S \subset \Omega: |S|=k\}} P(S)}{\sum_{m=k}^n (-1)^{m-k} \binom{n-k}{m-k} p_m}$$

$$= \binom{n}{k} \sum_{m=k}^n (-1)^{m-k} \binom{n-k}{m-k} p_m$$

$$= \binom{n}{k} \sum_{j=0}^{n-k} (-1)^j \binom{n-k}{j} p_{k+j}$$

Die Wahrscheinlichkeit, daß eine Stichprobe vom Umfang k gezogen wird, die vorgegebene, verschiedene Elemente $i_1, \dots, i_h$; $0 \leq h \leq k \leq n$ enthält, beträgt:

$$(1.1.12) \quad P(S_{i_1} = \dots = S_{i_h} = 1 \text{ und } \sum_{i=1}^n S_i = k)$$

$$= \frac{\sum_{\{S \subset \Omega: |S|=k \text{ und } \{i_1, \dots, i_h\} \subset S\}} P(S)}{\sum_{m=k}^n (-1)^{m-k} \binom{n-k}{m-k} p_m}$$

$$= \binom{n-h}{k-h} \sum_{m=k}^n (-1)^{m-k} \binom{n-k}{m-k} p_m$$

$$= \binom{n-h}{k-h} \sum_{j=0}^{n-k} (-1)^j \binom{n-k}{j} p_{k+j}$$

Nun werden die gebräuchlichsten Spezialfälle des symmetrischen Stichprobenverfahrens definiert.

Definition 1.1.4

Das symmetrische Stichprobenverfahren mit

$$(1.1.13) \quad p_m = \frac{\binom{n-m}{k-m}}{\binom{n}{k}} = \frac{\binom{k}{m}}{\binom{n}{m}} \quad (m=1, \dots, n; 1 \leq k \leq n)$$

heißt k -maliges Ziehen ohne Zurücklegen, wobei für $k < m$ die übliche Festlegung $\binom{k}{m} = 0$ bzw. $\binom{n-m}{k-m} = 0$ gelte (vgl. Feller (1962) S. 49). Da diese Stichprobenverfahren eine besondere Rolle spielen, bezeichnen wir sie mit $p^{(k)}$ und die in Gl. (1.1.13) definierten p_m mit $p_m^{(k)}$.

Der Stichprobenumfang k ist hier fest vorgegeben; in der Anwendung werden bei diesem Verfahren k -mal Elemente aus der Grundgesamtheit gezogen, wobei die schon ausgewählten Elemente im nächsten Zug nicht mehr gezogen werden können und in jedem Zug alle zur Wahl stehenden Elemente die gleiche Wahrscheinlichkeit haben, in die Stichprobe zu gelangen. Im ersten Zug wird also ein Element $i \in \Omega$ mit Wahrscheinlichkeit $\frac{1}{n}$ gezogen, im nächsten Zug wird aus den übrigen Elementen ein Element j mit Wahrscheinlichkeit $\frac{1}{n-1}$ gezogen usw..

Definition 1.1.5

Das symmetrisches Stichprobenverfahren mit

$$(1.1.14) \quad p_m = \sum_{j=0}^m (-1)^j \binom{m}{j} \left(1 - \frac{j}{n}\right)^{\ell} \quad (m=1, \dots, n; \quad 1 \leq \ell \leq n)$$

heißt ℓ -maliges Ziehen mit Zurücklegen.

Im Gegensatz zum vorigen Stichprobenverfahren stehen hier schon ausgewählte Elemente im nächsten Zug wieder zur Wahl, sie werden "zurückgelegt"; im ersten Zug wird also ein Element $i \in \Omega$ mit Wahrscheinlichkeit $\frac{1}{n}$ ausgewählt, im zweiten Zug wird unabhängig vom Ausgang des 1. Zuges wieder ein Element $j \in \Omega$ mit Wahrscheinlichkeit $\frac{1}{n}$ ausgewählt usw. Der Stichprobenumfang, d.h. die Anzahl der verschiedenen gezogenen Elemente, ist zufällig, da ein Element öfters gezogen werden kann; er kann Werte zwischen 1 und ℓ annehmen. Zur Ableitung der Formel für p_m aus der beschriebenen Ziehungsvorschrift siehe Feller (1962, S. 101).

Definition 1.1.6

Ein symmetrisches Stichprobenverfahren mit

$$(1.1.15) \quad p_m = p^m \quad (m=1, \dots, n; \quad 0 \leq p \leq 1)$$

heißt Bernoulli Stichprobenverfahren mit Auswahlwahrscheinlichkeit p .

Hier wird für jedes Element der Grundgesamtheit unabhängig voneinander ein Zufallsexperiment durchgeführt, das mit Wahrscheinlichkeit p ein Element für die Stichprobe auswählt. Die Stichprobengröße K ist eine binomialverteilte Zufallsvariable. Da hier

$$(1.1.16) \quad P(\{\emptyset\}) = (1-p)^n > 0 \text{ für } p < 1 \quad \text{gilt,}$$

kann es sinnvoll sein, gemäß der anschließenden Bemerkung 1.1.7 aus der Bernoulli Stichprobe Verfahren zu konstruieren, die sicherstellen, daß mindestens ein Element für die Stichprobe ausgewählt wird.

Bemerkung 1.1.7

In der Definition der symmetrischen Stichprobe war nicht ausgeschlossen, mit positiver Wahrscheinlichkeit die leere Menge als Stichprobe zu erhalten. Wenn der Fall $S=\emptyset$ auftritt, wird es sicherlich kein Anwender dabei belassen, sondern z.B. das Auswahlverfahren noch einmal anwenden. Außerdem sind in Abschnitt 1.3 und 2.2 die wichtigsten Schätzfunktionen nur für Stichprobenumfänge $K \geq 1$ bzw. $K \geq 2$ definiert.

Die oben angesprochene naheliegende Möglichkeit, aus einem beliebigen symmetrischen Stichprobenverfahren ebenfalls symmetrische Stichprobenverfahren P' bzw. P'' mit Stichprobenumfängen $K \geq 1$ bzw. $K \geq 2$ zu erzeugen, besteht in der Wiederholung des Auswahlverfahrens P bis der Stichprobenumfang genügend groß ist. P' und P'' werden dann durch bedingte Wahrscheinlichkeitsverteilungen definiert und es gilt:

$$(1.1.17) \quad P'(S) = P(S | K \geq 1) = \frac{P(S)}{1 - P(\{\emptyset\})}$$

für $|S| \geq 1$ bzw.:

$$(1.1.18) \quad P''(S) = P(S | K \geq 2) = \frac{P(S)}{1 - P(\{\emptyset\}) - nP(S^{(1)})}$$

für $|S| \geq 2$ und $|S^{(1)}| = 1$.

Satz 1.1.8

Ein Stichprobenverfahren P ist genau dann symmetrisch, wenn es $\alpha_k \geq 0$ mit $\sum_{k=0}^n \alpha_k = 1$ gibt, so daß

$$(1.1.19) \quad P = \sum_{k=0}^n \alpha_k P^{(k)},$$

wobei $P^{(0)}(S) := \begin{cases} 1 & \text{für } S=\emptyset \\ 0 & \text{sonst} \end{cases}$ sei und $P^{(k)}$ für $k \geq 1$

in Def. 1.1.4 festgelegt wurde.

Für die Wahrscheinlichkeiten $p_m, m=1, \dots, n$ und für α_k gilt:

$$(1.1.20) \quad p_m = \sum_{k=0}^n \alpha_k p_m^{(k)}$$

$$= \sum_{k=m}^n \alpha_k p_m^{(k)} \quad \text{mit } p_m^{(0)} := 0$$

$$(1.1.21) \quad \alpha_k = P\left(\sum_{i=1}^n S_i = k\right) \quad (k=0, \dots, n)$$

Beweis

Die Äquivalenz der Gleichungen (1.1.19) und (1.1.21) für symmetrische Stichprobenverfahren folgt aus der Definition der p_m bzw. aus Satz 1.1.2 (iii). Aus der Symmetrie der Stichprobenverfahren $P^{(k)}$ folgt sofort die Symmetrie des

Verfahrens $P = \sum_{k=0}^n \alpha_k P^{(k)}$; ein durch Gl. (1.1.19) gegebenes Stichprobenverfahren ist also symmetrisch.

Es bleibt zu zeigen, daß für jedes symmetrische Stichprobenverfahren Gleichung (1.1.19) oder (1.1.20) mit $\alpha_k = P(\sum_{i=1}^n S_i = k)$ gilt:

$$(1.1.22) \quad p_m = P(\Omega_m cS)$$

$$\begin{aligned} &= \sum_{k=0}^n \frac{P(\Omega_m cS \text{ und } \sum_{i=1}^n S_i = k)}{P(\sum_{i=1}^n S_i = k)} P(\sum_{i=1}^n S_i = k) \\ &= \sum_{k=0}^n \frac{\binom{n-m}{k-m} \sum_{j=0}^{n-k} (-1)^j \binom{n-k}{j} p_{k+j}}{\binom{n}{k} \sum_{j=0}^{n-k} (-1)^j \binom{n-k}{j} p_{k+j}} P(\sum_{i=1}^n S_i = k) \end{aligned}$$

(siehe Bemerkung 1.1.3)

$$\begin{aligned} &= \sum_{k=0}^n \frac{\binom{n-m}{k-m}}{\binom{n}{k}} \alpha_k = \sum_{k=0}^n \alpha_k p_m^{(k)} \\ &= \sum_{k=m}^n \alpha_k p_m^{(k)}, \text{ da } p_m^{(k)} = 0 \text{ für } k < m. \end{aligned}$$

Aus dem Gleichungssystem (1.1.20) kann man außerdem folgern, daß die $\alpha_k, k=1, \dots, n$ durch die $p_m, m=1, \dots, n$ eindeutig bestimmt werden und α_0 ist dann eindeutig durch $\alpha_0 = 1 - \sum_{k=1}^n \alpha_k$ festgelegt.

Bemerkung 1.1.9

Da für feste Stichprobengröße k_0 gilt: $\alpha_k = \begin{cases} 1 & \text{für } k=k_0 \\ 0 & \text{sonst} \end{cases}$

gibt es nach Satz 1.1.8 in diesem Fall genau ein symmetrisches Stichprobenverfahren, das k_0 -malige Ziehen ohne Zurücklegen.

Das bedingte Stichprobenverfahren $P^*(S) = P(S \mid \sum_{i=1}^n S_i = k_0)$ ist bei symmetrischen Stichprobenverfahren unabhängig von P auch das k_0 -malige Ziehen ohne Zurücklegen, da nach Bemerkung 1.1.3 gilt:

$$(1.1.23) \quad p_m^* = P(\Omega_m cS \mid \sum_{i=1}^n S_i = k_0) \\ = \frac{P(\Omega_m cS \text{ und } \sum_{i=1}^n S_i = k_0)}{P(\sum_{i=1}^n S_i = k_0)} = \frac{\binom{n-m}{k_0-m}}{\binom{n}{k_0}}$$

Bemerkung 1.1.10

Für ℓ -maliges Ziehen mit Zurücklegen gilt

$$(1.1.24) \quad \alpha_k = P(\sum_{i=1}^n S_i = k) = \binom{n}{k} \sum_{j=0}^k (-1)^j \binom{k}{j} \left(\frac{k-j}{n}\right)^\ell.$$

Da die Ableitung dieser Formel aus Gl. (1.1.11) und (1.1.14) sehr viel Raum in Anspruch nähme, sei hier nur auf Feller (1962), Bd. 1, S. 91,92 oder S. 58 verwiesen.

Bemerkung 1.1.11

Für Bernoulli Stichprobenverfahren mit Auswahlwahrscheinlichkeit p gilt:

$$\begin{aligned}
 (1.1.25) \quad \alpha_k &= \sum_{j=0}^{n-k} \binom{n}{k} (-1)^j \binom{n-k}{j} p^{k+j} \\
 &= \binom{n}{k} p^k \sum_{j=0}^{n-k} (-1)^j \binom{n-k}{j} p^j = \binom{n}{k} p^k (1-p)^{n-k}.
 \end{aligned}$$

Für die in Bemerkung 1.1.7 angesprochenen bedingten Stichprobenverfahren gilt dann

$$(1.1.26) \quad \alpha'_k = P\left(\sum_{i=1}^n S_i = k \mid \sum_{i=1}^n S_i \geq 1\right) = \frac{\alpha_k}{1 - \alpha_0}, \quad k=1, \dots, n \quad \text{und}$$

$$(1.1.27) \quad \alpha''_k = P\left(\sum_{i=1}^n S_i = k \mid \sum_{i=1}^n S_i \geq 2\right) = \frac{\alpha_k}{1 - \alpha_0 - \alpha_1}, \quad k=2, \dots, n.$$

Bemerkung 1.1.12

Nach Bemerkung 1.1.9 gilt für beliebige Funktionen f und m verschiedene Elemente $i_1, \dots, i_m$; $m=1, \dots, n$

$$\begin{aligned}
 (1.1.28) \quad E(f(K) S_{i_1} \dots S_{i_m}) &= E(f(K) E(S_{i_1} \dots S_{i_m} \mid K)) \\
 &= E\left(f(K) \frac{\binom{n-m}{K-m}}{\binom{n}{K}}\right) \\
 &= E\left(f(K) \frac{K(K-1) \dots (K-m+1)}{n(n-1) \dots (n-m+1)}\right)
 \end{aligned}$$

In den Beweisen von Satz 1.3.3 und 2.2.1 wird diese Gleichung eine wesentliche Rolle spielen.

Bemerkung 1.1.13

Abschließend wollen wir Parameter der Verteilung von K mittels p_m ; $m=1, \dots, n$ ausdrücken. Bekanntlich gilt:

$$(1.1.29) \quad E(K) = np_1$$

$$(1.1.30) \quad \text{var}(K) = np_1(1-np_1) + n(n-1)p_2$$

Etwas allgemeiner läßt sich der Erwartungswert von Produkten der Form $K(K-1)\dots(K-m)$ mit $m=0, \dots, n-1$ einfach angeben:

$$\begin{aligned} (1.1.31) \quad E(K(K-1)\dots(K-m)) &= \sum_{i_0=1}^n \sum_{\substack{i_1=1 \\ i_1 \neq i_0}}^n \dots \sum_{\substack{i_m=1 \\ i_m \neq i_j \\ \text{für } j < m}}^n S_{i_0} S_{i_1} \dots S_{i_m} \\ &= n(n-1)\dots(n-m)p_{m+1} \end{aligned}$$

1.2

Grundbegriffe der Schätztheorie

Zu jedem Element i der Grundgesamtheit $\Omega = \{1, 2, \dots, n\}$ seien Zufallsvariablen X_i definiert und der Vektor $X = (X_1, \dots, X_n)$ habe eine unbekannte Verteilung W aus einer Menge $\mathcal{W}$ von Verteilungen. Meistens ist $\mathcal{W}$ nicht die Menge aller n -dimensionalen Verteilungen, sondern eine Teilmenge, da man z.B. durch die Annahme der Unabhängigkeit der X_i oder durch die Annahmen, daß die X_i nur positive oder nur die Werte 0 und 1 haben können, Verteilungen aus $\mathcal{W}$ ausschließt. Geschätzt werden sollen ein oder mehrere Parameter der Verteilung W , in unserem Fall der Erwartungswert $\theta = \theta(W)$ der Summe der X_i

$$(1.2.1) \quad \theta = E \left(\sum_{i=1}^n X_i \right)$$

aus den Realisierungen der X_i in einer Stichprobe S . Eine Schätzfunktion $\hat{\theta} = \hat{\theta}(X, S)$ ordnet jeder möglichen Realisierung von X und jeder möglichen Stichprobe S einen Schätzwert für θ zu, ihre Verteilung wird durch W und das Stichprobenverfahren P bestimmt; einen Schätzwert (d.h. eine Realisierung von $\hat{\theta}$) bezeichnen wir mit $\hat{\theta}_O$.

In der Stichprobentheorie geht man normalerweise von festen, nicht zufälligen Werten $X_i \equiv x_i$ aus; anders ausgedrückt besteht die Familie $\mathcal{W}$ der möglichen Verteilungen aus n -dimensionalen Einpunktverteilungen. Die Zufallsschwankungen der Schätzfunktion $\hat{\theta}$ werden dann nur durch das Stichprobenverfahren P verursacht.

In der Schätztheorie werden zu allen Elementen i der Grundgesamtheit Realisierungen der Zufallsvariablen X_i beobachtet, um daraus Parameter der Verteilung W zu schätzen.

Das einzige betrachtete Stichprobenverfahren wählt also Ω mit Wahrscheinlichkeit 1 aus und die Verteilung der Schätzfunktion $\hat{\theta}$ wird nur durch W bestimmt. Formal genügen allerdings auch andere Stichprobenverfahren diesem Modell, wenn man als $\mathcal{W}$ die Menge aller möglichen Verteilungen der Zufallsvariablen $(X_1 S_1, \dots, X_i S_i, \dots, X_n S_n)$ wählt, wobei S_i die Stichprobenindikatorvariablen sind; wir werden jedoch die Verteilungen der $X = (X_1, \dots, X_n)$ und das Stichprobenverfahren P getrennt betrachten.

Eine Schätzstrategie besteht nun aus der Wahl eines Stichprobenverfahrens P und einer Schätzfunktion $\hat{\theta}$; wir werden P als fest vorgegeben betrachten und nur Kriterien für die Wahl einer Schätzfunktion entwickeln. Die folgenden Definitionen von Güteeigenschaften von Schätzfunktionen gelten allerdings analog für Schätzstrategien.

Ein naheliegendes Kriterium für die Güte von Schätzfunktionen ist die Erwartungstreue, die Schätzfunktion soll "im Mittel" die richtigen Werte liefern:

Definition 1.2.1

Eine Schätzfunktion heißt erwartungstreu für die Parameter $\theta (= \theta(W))$, wenn gilt:

$$(1.2.2) \quad E_W(\hat{\theta}) = \theta \quad \text{für alle } W \in \mathcal{W},$$

wobei $E_W(\cdot)$ Bildung des Erwartungswertes bezüglich der von W und P erzeugten Verteilung von $\hat{\theta}$ bedeutet.

Die Erwartungstreue als einziges Kriterium ist nicht ausreichend, man wird z.B. nach erwartungstreuen Schätzfunktionen mit möglichst kleiner Varianz suchen. Außerdem kann es Gründe geben, eine nicht erwartungstreue Schätzfunktion

(z.B. mit kleinem systematischem Fehler aber auch kleiner Varianz) einer erwartungstreuen (z.B. mit sehr großer Varianz) vorzuziehen. Aus diesen Gründen erweist sich ein anderes, der Spieltheorie entlehntes Konzept zur Schaffung von Gütekriterien sinnvoller (vgl. Ferguson (1967), Witting (1966), Wald (1950), Blackwell, Girshick (1954)). Ausgangspunkt ist der Begriff der Verlustfunktion $L(\hat{\theta}_0, \theta)$, die jedem möglichen Schätzwert $\hat{\theta}_0$ und jedem möglichen Parameter θ eine positive reelle Zahl als den Verlust zuordnet, der entsteht, wenn der Parameter θ durch $\hat{\theta}_0$ geschätzt wird. Bei einem richtigen Schätzwert soll der Verlust 0 betragen, d.h. $L(\hat{\theta}_0, \theta) = 0$ für $\hat{\theta}_0 = \theta$. Diese Verlustfunktion kann man in der Praxis durch die Unkosten oder allgemeiner durch eine Bewertung der Konsequenzen bestimmen, die durch eine falsche Schätzung entstehen. Da es oft zu kompliziert oder unmöglich ist, diese Konsequenzen quantitativ zu bewerten und da die theoretische Statistik nicht auf alle möglichen speziellen Verlustfunktionen eingehen kann, wird man sich auf analytisch möglichst einfache Verlustfunktionen beschränken, die eine Bewertung des Schätzfehlers durch den Statistiker wiedergeben und zumindest näherungsweise oft der Praxis entsprechen; durchgesetzt hat sich hierbei in der Schätztheorie die Gaußsche (quadratische) Verlustfunktion für Parameter $\theta \in \mathbb{R}$:

$$(1.2.3) \quad L(\hat{\theta}_0, \theta) = (\hat{\theta}_0 - \theta)^2,$$

oder allgemeiner

$$(1.2.4) \quad L(\hat{\theta}_0, \theta) = f(\theta) (\hat{\theta}_0 - \theta)^2$$

mit positiver, reeller Funktion f .

Die in (1.2.3) definierte Verlustfunktion gewichtet Über- und Unterschätzungen gleich, große Abweichungen vom wahren Wert werden überproportional gewichtet.

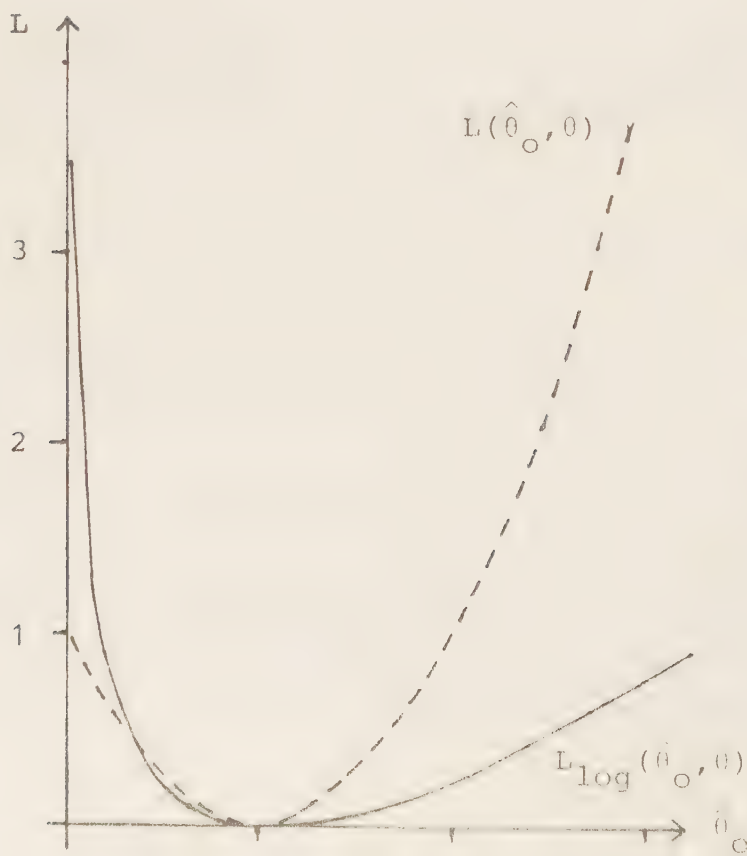
Falls $\theta > 0$ und $\hat{\theta}_0 > 0$ stets gilt, kann es sinnvoll sein, nicht wie in (1.2.3) die absoluten Abweichungen des Schätzers vom wahren Wert $|\hat{\theta}_0 - \theta|$, sondern die relativen Abweichungen $\left(\frac{\hat{\theta}_0 - \theta}{\theta}\right) = \frac{\hat{\theta}_0}{\theta} - 1$ zu bewerten und somit Verlustfunktionen zu betrachten, die Funktionen des Quotienten $\frac{\hat{\theta}_0}{\theta}$ sind. Die Gaußsche Verlustfunktion mit $f(\theta) = \frac{1}{\theta^2}$ ist eine derartige Verlustfunktion:

$$(1.2.5) \quad L(\hat{\theta}_0, \theta) = \frac{1}{\theta^2} (\hat{\theta}_0 - \theta)^2 = \left(\frac{\hat{\theta}_0}{\theta} - 1 \right)^2.$$

Eine andere Möglichkeit, die relativen Abweichungen zu gewichten, bietet die von Stenger (1979) behandelte logarithmische Verlustfunktion:

$$(1.2.6) \quad L_{\log}(\hat{\theta}_0, \theta) = g(\theta) \left(\frac{\hat{\theta}_0}{\theta} - \log \frac{\hat{\theta}_0}{\theta} - 1 \right),$$

wobei g eine positive, reelle Funktion und $\hat{\theta}_0, \theta > 0$ sind.



Die nebenstehende Zeichnung soll eine Vorstellung vom Verlauf der beiden letztgenannten Verlustfunktionen (für $g(\theta) \equiv 1$) und damit von der Bewertung der Schätzfehler geben.

Die Symmetrie der Gaußschen Verlustfunktion ist bei positiven Parametern nicht immer wünschenswert, z.B. ist der Verlust bei jeder Unterschätzung geringer als bei einer Überschätzung um das Doppelte (d.h. $\frac{\hat{\theta}}{\theta} = 2$). Die logarithmische Verlustfunktion bewertet große relative Unterschätzung ($\frac{\hat{\theta}}{\theta} \leq 0,3$) bedeutend mehr und Überschätzungen bedeutend weniger als die quadratische; sie ist allerdings nicht gut interpretierbar, weil man z.B. nicht explizit angeben kann, bei welcher Unterschätzung des wahren Wertes die Verlustfunktion denselben Wert annimmt wie bei einer Überschätzung um das α -fache ($\frac{\hat{\theta}}{\theta} = \alpha > 1$). Der Hauptgrund für die Einführung der logarithmischen Verlustfunktion liegt darin, daß bei den hier behandelten Problemen unter Zugrundelegung dieser Verlustfunktion eine "beste" Schätzfunktion existiert (vgl. Abschnitt 1.4, 2.3 und Stenger (1979)), wobei der Begriff "beste Schätzfunktion" noch zu klären ist.

Wenn wir anstelle eines festen Schätzwertes $\hat{\theta}_0$ eine Schätzfunktion, also eine Zufallsvariable betrachten, liegt es nahe, die Erwartungswerte der Verlustfunktion bezüglich den von P und $W \in \mathcal{W}$ erzeugten Verteilungen zu bilden. Bei gegebener Verlustfunktion ordnet die Risikofunktion R dann jeder Schätzfunktion $\hat{\theta}$ und jeder möglichen Verteilung W der Werte den erwarteten Verlust zu, wenn bei Vorliegen von W der Parameter $\theta = \theta(W)$ dieser Verteilung durch $\hat{\theta}$ geschätzt wird:

$$(1.2.7) \quad R(\hat{\theta}, W) = E_W(L(\hat{\theta}, \theta))$$

Man wird nun Schätzfunktionen mit möglichst kleinen Risikofunktionen suchen. Falls nur erwartungstreue Schätzfunktionen und die Gaußsche Verlustfunktion (Gl. 1.2.3) betrachtet werden, gibt die Risikofunktion für jedes W und jede Schätzfunktion $\hat{\theta}$ die Varianz von $\hat{\theta}$ an.

Definition 1.2.2

- (i) Schätzfunktionen $\hat{\theta}$ und $\hat{\hat{\theta}}$ heißen äquivalent in $\mathcal{W}$ (bezüglich einer Verlustfunktion L), wenn gilt:

$$(1.2.8) \quad R(\hat{\theta}, W) = R(\hat{\hat{\theta}}, W) \quad \text{für alle } W \in \mathcal{W}.$$

- (ii) Eine Schätzfunktion $\hat{\theta}$ heißt gleichmäßig besser als $\hat{\hat{\theta}}$ in $\mathcal{W}$, wenn $\hat{\theta}$ und $\hat{\hat{\theta}}$ nicht äquivalent sind und wenn gilt:

$$(1.2.9) \quad R(\hat{\theta}, W) \leq R(\hat{\hat{\theta}}, W) \quad \text{für alle } W \in \mathcal{W}.$$

Da es oft zu schwer oder unmöglich ist, alle denkbaren Schätzfunktionen zu vergleichen, beschränkt man sich auf Teilmengen Ψ der Menge aller Schätzfunktionen. In der Schätztheorie betrachtet man meistens nur erwartungstreue Schätzfunktionen, in der Stichprobentheorie bei symmetrischen Stichprobenverfahren die linearen, symmetrischen Schätzfunktionen.

Definition 1.2.3

Eine Schätzfunktion $\hat{\theta}$ heißt linear und symmetrisch, wenn sie die Form hat:

$$(1.2.10) \quad \hat{\theta} = \beta(K) \sum_{i=1}^n X_i S_i,$$

wobei K der Stichprobenumfang, β eine reelle Funktion und S_i die Stichprobenindikatorvariablen sind.

Lineare, symmetrische Schätzer sind also Linearkombinationen der in der Stichprobe beobachteten Werte, wobei die Koeffizienten $\beta(K)$ dieser Linearkombination nur vom Stichprobenumfang abhängen. Diese Schätzfunktionen werden

vollständig durch den Vektor $(\beta(1), \beta(2), \dots, \beta(n))$ bestimmt. Die Einschränkung auf lineare, symmetrische Schätzfunktionen hat den Vorteil, daß die Risikofunktion für die hier behandelten Probleme nur noch von β und wenigen Parametern der Verteilung W abhängt (vgl. Abschnitt 1.3 und 2.2).

Wir definieren nun die wichtigsten Gütekriterien:

Definition 1.2.4

Gegeben sei eine Menge $\mathcal{W}$ von n -dimensionalen Wahrscheinlichkeitsverteilungen und eine Menge Ψ von Schätzfunktionen für einen Parameter $\theta = \theta(W)$.

- (i) Eine Schätzfunktion $\hat{\theta} \in \Psi$ heißt gleichmäßig beste Schätzfunktion bzgl. $\mathcal{W}$ und Ψ , wenn für alle $\hat{\theta} \in \Psi$ und $W \in \mathcal{W}$ gilt:

$$R(\hat{\theta}, W) \leq R(\hat{\theta}, W)$$

- (ii) Eine Schätzfunktion $\hat{\theta} \in \Psi$ heißt zulässig bzgl. $\mathcal{W}$ und Ψ , wenn es keine gleichmäßig bessere Schätzfunktion gibt.

- (iii) Eine Klasse von Schätzfunktionen $\Psi' \subset \Psi$ heißt vollständig in Ψ bzgl. $\mathcal{W}$, wenn es zu jeder Schätzfunktion $\hat{\theta} \in \Psi$ mit $\hat{\theta} \notin \Psi'$ eine Schätzfunktion $\hat{\theta} \in \Psi'$ gibt, die gleichmäßig besser ist.

Bemerkung 1.2.5

Gleichmäßig beste Schätzfunktionen sind stets zulässig und sie bilden die kleinste vollständige Klasse von Schätzfunktionen. Falls keine gleichmäßig beste Schätzfunktion existiert, wird man nach möglichst kleinen vollständigen Klassen von Schätzfunktionen suchen. Alle vollständigen Klassen enthalten die Menge der zulässigen Schätzfunktionen und falls eine

minimale vollständige Klasse existiert (d.h. eine vollständige Klasse, die Teilmenge aller anderen vollständigen Klassen ist), ist es die Menge der zulässigen Schätzfunktionen (vgl. Ferguson (1967) S. 55/56, oder Wald (1950) S. 54). Allerdings existiert nicht immer eine minimale vollständige Klasse von Schätzfunktionen bzw. die Menge der zulässigen Schätzfunktionen ist nicht immer vollständig; eine Bedingung an $\mathcal{W}$, Ψ und L dafür, daß die Menge der zulässigen Schätzfunktionen die minimale vollständige Klasse ist, findet sich z.B. in Wald (1950) S. 54 oder Blackwell, Girshick (1954) S. 141).

Bemerkung 1.2.6

Die in Def. 1.2.4 eingeführten Eigenschaften hängen von der Wahl von $\mathcal{W}$, Ψ und der Verlustfunktion L ab, z.B. kann eine erwartungstreue, in der Klasse der linearen, symmetrischen Schätzfunktionen unzulässige Schätzfunktion in der Klasse der erwartungstreuen und linearen, symmetrischen Schätzfunktion zulässig sein. Für zwei Klassen von Schätzfunktionen Ψ' und Ψ mit $\Psi' \subset \Psi$ gilt: falls $\hat{\theta} \in \Psi$ und $\hat{\theta}$ zulässig in Ψ ist, ist $\hat{\theta}$ auch zulässig in Ψ' . Ebenfalls direkt aus der Definition ergibt sich folgende Aussage zur Menge der Verteilungen $\mathcal{W}$: falls $\hat{\theta}$ bzgl. einer Teilmenge $\mathcal{W}' \subset \mathcal{W}$ die einzige gleichmäßig beste Schätzfunktion ist, ist sie bzgl. $\mathcal{W}$ zulässig.

Bemerkung 1.2.7

Da für beliebige Verlustfunktionen L und reelle Funktionen f mit $f(\theta) > 0$ stets gilt:

$$(1.2.11) \quad L(\hat{\theta}_0, \theta) = L(\hat{\hat{\theta}}_0, \theta) \text{ ist äquivalent zu}$$

$$f(\theta)L(\hat{\theta}_0, \theta) = f(\theta)L(\hat{\hat{\theta}}_0, \theta)$$

und

(1.2.12) $L(\hat{\theta}_0, \theta) \leq L(\hat{\hat{\theta}}_0, \theta)$ ist äquivalent zu

$$f(\theta)L(\hat{\theta}_0, \theta) \leq f(\theta)L(\hat{\hat{\theta}}_0, \theta),$$

sind die in Def. 1.2.2 und 1.2.4 gebrachten Eigenschaften unabhängig von der Wahl der Funktion f bei der allgemeinen Gaußschen Verlustfunktion (1.2.4) und der Funktion g bei der logarithmischen Verlustfunktion (1.2.6), sofern $f(\theta) > 0$ und $g(\theta) > 0$ stets gilt. Wir können also ohne Einschränkung der Allgemeinheit $f(\theta) = g(\theta) \equiv 1$ setzen.

Bemerkung 1.2.8

Oft will man nicht den Erwartungswert θ der Summe der X_i (Gl. 1.2.1) schätzen, sondern z.B. den Erwartungswert des arithmetischen Mittels $\frac{1}{n}\theta$ oder allgemein Parameter $\alpha_0\theta$ mit $\alpha_0 \in \mathbb{R}$, $\alpha_0 \neq 0$. Bei Zugrundelegung der Gaußschen oder logarithmischen Verlustfunktion gilt für die Schätzfunktionen $\alpha_0\hat{\theta}$:

$$(1.2.13) \quad E(\alpha_0\hat{\theta}) = \alpha_0 E(\hat{\theta}) \quad \text{und}$$

$$(1.2.14) \quad R(\alpha_0\hat{\theta}, \alpha_0\theta) = f(\alpha_0)R(\hat{\theta}, \theta)$$

mit $f(\alpha_0) = \alpha_0^2$ bei der Gaußschen und $f(\alpha_0) = 1$ bei der logarithmischen Verlustfunktion.

Falls nun eine der hier definierten Güteeigenschaften für $\hat{\theta}$ als Schätzfunktion von θ in der Menge der Schätzfunktionen Ψ zutrifft, trifft sie auch für $\alpha_0\hat{\theta}$ als Schätzfunktion von $\alpha_0\theta$ in $\Psi' = \{\alpha\hat{\theta} : \hat{\theta} \in \Psi\}$ zu (vgl. Def. 1.2.1, 1.2.2, 1.2.4 und Bem. 1.2.7). Wenn wir nun z.B. eine zulässige Schätzfunktion $\hat{\theta}$ für die Summe der Werte haben, ist $\frac{1}{n}\hat{\theta}$ zulässige Schätzfunktion (in der entsprechenden Klasse) für das arithmetische Mittel der Werte.

1.3 Die Schätzung des Erwartungswertes der Summe der Zufallsvariablen

Im folgenden werden die Erwartungswerte von beliebigen linearen, symmetrischen Schätzfunktionen für θ und die Werte der Risikofunktion bei Gaußscher Verlustfunktion berechnet. Bei der Definition weiterer Parameter und bei der Erläuterung der Ergebnisse werden die Besonderheiten der Modellvorstellung einer Superpopulation betont. Weiterhin gehen wir auf die beiden gebräuchlichsten Schätzfunktionen ein; das n -fache des arithmetischen Mittels der beobachteten Werte und die Horvitz-Thompson-Schätzfunktionen.

Die allgemeinen Überlegungen des letzten Abschnittes treffen sowohl auf einzelnen Untersuchungseinheiten zugeordnete Zufallsvariablen X_i , als auch auf Zufallsvariablen zu, die Paaren von Untersuchungseinheiten zugeordnet sind. Um im weiteren Verlauf dieses Kapitels zu kennzeichnen, daß sich die Aussagen auf Merkmale der einzelnen Untersuchungseinheiten beziehen, verwenden wir bei Parametern, Schätzfunktionen usw. als ersten Index 1. Die Summe der X_i bezeichnen wir $T_1 := \sum_{i=1}^n X_i$. Geschätzt werden soll der Erwartungswert θ_1 dieser Summe durch eine lineare, symmetrische Schätzfunktion $\hat{\theta}_1$ aus einer symmetrischen Stichprobe. Zur Angabe von Eigenschaften dieser Schätzfunktionen ist bei Superpopulationen eine Klärung der Beziehung zwischen dem Stichprobenverfahren und den Zufallsvariablen X_i notwendig.

Voraussetzung 1.3.1

Die Familie von Zufallsvariablen $\{X_i: i \in \Omega\}$ sei von der Menge der Stichprobenindikatorvariablen $\{S_i: i \in \Omega\}$ stochastisch unabhängig.

Die durch Funktionen $f(X_1, \dots, X_n)$ und $g(S_1, \dots, S_n)$ gebildeten Zufallsvariablen sind damit auch stochastisch unabhängig (vgl. Krickeberg (1962), S. 76) und speziell gilt:

$$\begin{aligned} (1.3.1) \quad E(f(X_1, \dots, X_n)g(S_1, \dots, S_n)) \\ = E(f(X_1, \dots, X_n))E(g(S_1, \dots, S_n)) \end{aligned}$$

Die Verteilung W von $(X_1, \dots, X_n)$ und ein symmetrisches Stichprobenverfahren P seien in diesem Abschnitt fest vorgegeben. Wahrscheinlichkeiten, Erwartungswerte und andere Parameter beziehen sich auf die von W und P erzeugte Wahrscheinlichkeitsverteilung der betrachteten Zufallsvariablen.

Bemerkung 1.3.2

Es sei darauf hingewiesen, daß bei ℓ -maligem Ziehen mit Zurücklegen (vgl. Def. 1.1.5) eine gängige Schätzfunktion nicht unter den hier gebrauchten Begriff der linearen, symmetrischen Schätzfunktion fällt. Wenn $x_{(j)}$ den im j -ten Zug beobachteten Wert bezeichnet, wird oft folgende Schätzung für das Mittel $\mu_1 = \frac{1}{n}\theta_1$ betrachtet (vgl. Stenger (1971) S. 66)

$$(1.3.2) \quad \hat{\mu} = \frac{1}{\ell} \sum_{j=1}^{\ell} x_{(j)}$$

Weil sich der Index (j) auf die Ziehung und nicht auf die Untersuchungseinheit bezieht, können die im j -ten und m -ten Zug beobachteten Werte zur selben Untersuchungseinheit $i \in \Omega$ gehören. Bei festen zugrundeliegenden Variablenwerten werden

dann die Werte $x_i, i \in S$ verschieden gewichtet, je nachdem wie oft sie gezogen werden.

Satz 1.3.3

Für $\hat{\theta}_1 = \beta(K) \sum_{i=1}^n x_i s_i$ gilt:

$$(1.3.3) \quad E(\hat{\theta}_1) = \theta_1 E\left(\beta(K) \frac{K}{n}\right) \quad \text{und}$$

$$(1.3.4) \quad E(\hat{\theta}_1 - \theta_1)^2 = a_{11} E\left(\beta^2(K) \frac{K}{n}\right) + a_{12} E\left(\beta^2(K) \frac{K(K-1)}{n(n-1)}\right) \\ - \theta_1^2 (2E\left(\beta(K) \frac{K}{n}\right) - 1)$$

mit

$$(1.3.5) \quad a_{11} := \sum_{i=1}^n E(x_i^2) \quad \text{und}$$

$$(1.3.6) \quad a_{12} := \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n E(x_i x_j) = 2 \sum_{i=1}^n \sum_{j>i}^n E(x_i x_j)$$

Beweis

Nach Gleichung (1.1.28) und wegen der Unabhängigkeit der Werte von dem Stichprobenverfahren (vgl. Gl. 1.3.1) gilt:

$$\begin{aligned}
 (1.3.7) \quad E(\hat{\theta}_1) &= E\left(\beta(K) \sum_{i=1}^n X_i S_i\right) \\
 &= \sum_{i=1}^n E(X_i) E(\beta(K) S_i) \\
 &= \sum_{i=1}^n E(X_i) E\left(\beta(K) \frac{K}{n}\right) \\
 &= \theta_1 E\left(\beta(K) \frac{K}{n}\right) \quad \text{und weiterhin:}
 \end{aligned}$$

$$\begin{aligned}
 (1.3.8) \quad E(\hat{\theta}_1 - \theta_1)^2 &= E(\hat{\theta}_1^2) - 2\theta_1 E(\hat{\theta}_1) + \theta_1^2 \\
 &= \sum_{i=1}^n \sum_{j=1}^n E(X_i X_j) E(\beta^2(K) S_i S_j) \\
 &\quad - \theta_1^2 \left(2E\left(\beta(K) \frac{K}{n}\right) - 1\right) \\
 &= a_{11} E\left(\beta^2(K) \frac{K}{n}\right) + a_{12} E\left(\beta^2(K) \frac{K(K-1)}{n(n-1)}\right) \\
 &\quad - \theta_1^2 \left(2E\left(\beta(K) \frac{K}{n}\right) - 1\right)
 \end{aligned}$$

Bemerkung 1.3.4

Lineare und symmetrische Schätzfunktionen sind nach Satz 1.3.3 also genau dann erwartungstreu, wenn

$$(1.3.9) \quad E\left(\beta(K) \frac{K}{n}\right) = 1 \quad \text{gilt.}$$

Insbesondere werden wir die erwartungstreuen Schätzfunktionen mit $\beta(K) = \frac{n}{K}$ und $\beta(K) = \frac{n}{E(K)}$ betrachten; wir erhalten also im ersten Fall das n -fache des arithmetischen Mittels der Stichprobenwerte als erwartungstreue Schätzfunktion:

$$(1.3.10) \quad \theta_1^* := \frac{n}{K} \sum_{i=1}^n X_i S_i$$

und im zweiten Fall die Horvitz-Thompson Schätzfunktion:

$$(1.3.11) \quad \tilde{\theta}_1 := \frac{n}{E(K)} \sum_{i=1}^n X_i S_i = \frac{1}{p_1} \sum_{i=1}^n X_i S_i.$$

Bei fester Stichprobengröße sind beide Schätzfunktionen identisch. Es sei noch betont, daß θ_1^* nur für Stichprobenverfahren mit $\alpha_0 = P(K=0) = 0$ definiert ist. Für die quadratische Abweichung der Werte vom wahren Parameter gilt:

$$(1.3.12) \quad E(\theta_1^* - \theta_1)^2 = \text{var}(\theta_1^*) = a_{11} n E\left(\frac{1}{K}\right) + a_{12} \frac{n}{n-1} \left(1 - E\left(\frac{1}{K}\right)\right) - \theta_1^2$$

$$(1.3.13) \quad E(\tilde{\theta}_1 - \theta_1)^2 = \text{var}(\theta_1^*)$$

$$= a_{11} n \frac{1}{E(K)} + a_{12} \frac{n}{n-1} \frac{E(K^2) - E(K)}{E^2(K)} - \theta_1^2$$

$$= a_{11} \frac{1}{p_1} + a_{12} \frac{p_2}{p_1^2} - \theta_1^2$$

O. Frank (1969, 1971, 1977a, 1978b) betrachtet nur Horvitz-Thompson-Schätzfunktionen bzw. Schätzfunktionen bei fest vorgegebener Stichprobengröße für nicht zufällige Variablenwerte und erhält für die Varianz das in Gl. (1.3.13) angegebene Ergebnis. Ein Vergleich der Güte von $\tilde{\theta}_1$ mit den anderen linearen, symmetrischen Schätzfunktionen, insbesondere mit θ_1^* wird später durchgeführt.

Da die Parameter a_{11} und a_{12} in Gl. (1.3.4) relativ unanschaulich und daher schlecht zu interpretieren sind, sollen im folgenden Satz diese Parameter durch Maße für die Streuung der Variablenwerte in der Grundgesamtheit ersetzt werden.

In der Stichprobentheorie, wo die Variablenwerte x_i normalerweise nicht zufällig sind, wird die Streuung durch

$$(1.3.14) \quad \sigma_1^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \frac{1}{n} \sum_{i=1}^n x_i)^2$$

definiert, wobei oft statt des Faktors $\frac{1}{n}$ auch $\frac{1}{n-1}$ genommen wird. Da anstelle von Werten x_i hier im allgemeinen von Zufallsvariablen X_i ausgegangen wird, würde (1.3.14) eine Zufallsvariable und keinen Parameter definieren. Wir haben also einen einfachen und gut zu interpretierenden Parameter für die Streuung der Werte so zu definieren, daß Gl. (1.3.14) ein Spezialfall dieser Definition für den Fall $\text{var}(x_i)=0$ für alle $i \in \Omega$ ist. Eine naheliegende Lösung ist:

$$(1.3.15) \quad \begin{aligned} \sigma_1^2 &:= E \left(\frac{1}{n} \sum_{i=1}^n (X_i - \frac{1}{n} \sum_{i=1}^n X_i)^2 \right) \\ &= \frac{1}{n} a_{11} - \frac{1}{n^2} E(T_1^2) = \frac{n-1}{n^2} a_{11} - \frac{1}{n^2} a_{12} . \end{aligned}$$

Das letzte Gleichheitszeichen folgt hierbei aus:

$$(1.3.16) \quad E(T_1^2) = a_{11} + a_{12} = \text{var}(T_1) + \theta_1^2$$

In σ_1^2 gehen sowohl Unterschiede der Erwartungswerte der X_i als auch ihre Varianzen und Kovarianzen ein. Es erscheint daher sinnvoll, zusätzlich die Streuung von $E(X_i), i=1, \dots, n$ als Parameter einzuführen und zu σ_1^2 in Beziehung zu setzen.

$$\begin{aligned}
 (1.3.17) \quad \sigma_{11}^2 &:= \frac{1}{n} \sum_{i=1}^n (E(X_i) - \frac{1}{n} \theta_1)^2 \\
 &= \frac{1}{n} \sum_{i=1}^n E^2(X_i) - \frac{1}{n^2} \theta_1^2 \\
 &= \frac{1}{n} a_{11} - \frac{1}{n} \sum_{i=1}^n \text{var}(X_i) - \frac{1}{n^2} E(T_1^2) + \frac{1}{n^2} \text{var}(T_1) \\
 &= \sigma_1^2 - \frac{n-1}{n^2} \sum_{i=1}^n \text{var}(X_i) + \frac{2}{n^2} \sum_{i=1}^n \sum_{j>i}^n \text{cov}(X_i, X_j)
 \end{aligned}$$

Falls die Variablenwerte nicht als Realisierungen von Zufallsvariablen aufgefaßt werden, stimmen σ_1^2 und σ_{11}^2 überein.

Satz 1.3.5

Es gilt:

$$\begin{aligned}
 (1.3.18) \quad E(\hat{\theta}_1 - \theta_1)^2 &= \sigma_1^2 E(\beta^2(K) \frac{K(n-K)}{n-1}) \\
 &\quad + \text{var}(T_1) E(\beta(K) \frac{K}{n})^2 \\
 &\quad + \theta_1^2 E(\beta(K) \frac{K}{n} - 1)^2 \\
 &= \sigma_{11}^2 E(\beta^2(K) \frac{K(n-K)}{n-1}) \\
 &\quad + \sum_{i=1}^n \text{var}(X_i) E(\beta^2(K) \frac{K}{n}) \\
 &\quad + 2 \sum_{i=1}^n \sum_{j>i}^n \text{cov}(X_i, X_j) E(\beta^2(K) \frac{K(K-1)}{n(n-1)}) \\
 &\quad + \theta_1^2 E(\beta(K) \frac{K}{n} - 1)^2
 \end{aligned}$$

Beweis

Die erste Gleichung erhält man durch Einsetzen in Gl. (1.3.4) von

$$(1.3.19) \quad a_{11} = \frac{1}{n} \text{var}(T_1) + \frac{1}{n} \theta^2 + n \sigma_1^2 \quad \text{und}$$

$$(1.3.20) \quad a_{12} = \frac{n-1}{n} \text{var}(T_1) + \frac{n-1}{n} \theta^2 + n \sigma_1^2 \quad .$$

Die zweite Gleichung erhält man aus der ersten durch Einsetzen der nach σ_1^2 aufgelösten Gl. (1.3.17) und durch Aufteilung von $\text{var}(T_1)$:

$$\begin{aligned} (1.3.21) \quad \text{var}(T_1) &= \sum_{i=1}^n \sum_{j=1}^n \text{cov}(X_i, X_j) \\ &= \sum_{i=1}^n \text{var}(X_i) + 2 \sum_{i=1}^n \sum_{j>i}^n \text{cov}(X_i, X_j) \quad . \end{aligned}$$

Bemerkung 1.3.6

Aus Gleichung (1.3.17) sieht man, daß σ_1^2 sich aus der Streuung der Mittelwerte σ_{11}^2 und den Varianzen und Kovarianzen der Variablen zusammensetzt, und daher unterscheidet σ_1^2 nicht, ob die Streuung der Werte durch unterschiedliche Mittelwerte $E(X_i)$ oder durch $\text{var}(X_i)$ verursacht wurde. Es ist also plausibel, daß man bei der Darstellung von $E(\hat{\theta}_1 - \theta_1)^2$ durch die Streuung der Mittelwerte σ_{11}^2 einen Parameter mehr als im ersten Teil von Gl. (1.3.18) oder in Gl. (1.3.4) braucht. Da man bei einem stochastischen Modell meistens Vorstellungen über die Varianzen und Kovarianzen der Werte hat, haben wir im zweiten Teil von Gl. (1.3.18) ihre Summen als weitere Parameter eingeführt.

Ergänzend sei noch gesagt, daß man anstelle der Summe der Kovarianzen aller Wertepaare die Summe der Kovarianzen der Werte zur Wertsumme T_1 als Parameter einführen kann, da gilt:

$$(1.3.22) \quad 2 \sum_{i=1}^n \sum_{j>i}^n \text{cov}(X_i, X_j) = \sum_{i=1}^n \text{cov}(X_i, T_1) - \sum_{i=1}^n \text{var}(X_i).$$

Bemerkung 1.3.7

Bei erwartungstreuen, linearen, symmetrischen Schätzfunktionen gilt für den Koeffizienten von θ_1^2 :

$$(1.3.23) \quad E\left(\beta(K) \frac{K}{n} - 1\right)^2 = \text{var}\left(\beta(K) \frac{K}{n}\right)$$

und speziell für θ_1^* beträgt dieser Faktor 0; d.h. der erwartete quadratische Fehler hängt nicht von der Größe des wahren θ_1 ab und es ergibt sich die einfache, für den Spezialfall $\text{var}(T_1)=0$ in der Stichprobentheorie bekannte Formel:

$$(1.3.24) \quad \text{var}(\theta_1^*) = \frac{n^3}{n-1} \sigma_1^2 \left(E\left(\frac{1}{K}\right) - \frac{1}{n}\right) + \text{var}(T_1)$$

Für die ebenfalls erwartungstreuen Horvitz-Thompson-Schätzfunktionen gilt:

$$(1.3.25) \quad \text{var}(\tilde{\theta}_1) = \sigma_1^2 \frac{n(p_1 - p_2)}{p_1^2} + \text{var}(T_1) \frac{p_1 - (n-1)p_2}{np_1^2}$$

$$+ \theta_1^2 \frac{p_1(1-np_1) + (n-1)p_2}{np_1^2}$$

Bemerkung 1.3.8

In dieser Bemerkung soll auf spezielle Verteilungsannahmen eingegangen werden. Falls man von festen Variablenwerten in der Grundgesamtheit ausgeht, betragen alle Varianzen und Kovarianzen in Gl. (1.3.18) 0; für unabhängig voneinander verteilte Variablenwerte betragen die Kovarianzen 0 bzw. es gilt $\text{var}(T_1) = \sum_{i=1}^n \text{var}(X_i)$. Falls alle X_i denselben Mittelwert haben, speziell also bei identisch verteilten Variablen, gilt $\sigma_{11}^2 = 0$. Für identisch verteilte, unabhängige X_i mit $E(X_i) = w_1$ und $\text{var}(X_i) = v_1$ erhalten wir den Standardfall der Statistik mit dem einzigen Unterschied, daß der Stichprobenumfang variabel ist. Es gilt:

$$(1.3.26) \quad E(\hat{\theta}_1 - \theta_1)^2 = v_1 E(\beta^2(K)K) + w_1 E(\beta(K)K-n)^2$$

Bei Schätzung von $\mu_1 = \frac{1}{n}\theta_1$ durch $\mu_1^* = \frac{1}{n}\theta_1^* = \frac{1}{K} \sum_{i=1}^n X_i S_i$ erhält man dann das bekannte Ergebnis:

$$(1.3.27) \quad \text{var}(\mu_1^*) = v_1 E\left(\frac{1}{K}\right).$$

Bemerkung 1.3.9

Hier sollen die Konsequenzen betrachtet werden, die beim Übergang von der in der Stichprobentheorie üblichen Auffassung der Werte als feste Zahlen zu der Vorstellung der Werte als Realisierung von Zufallsvariablen entstehen. Aus Gl. (1.3.3) sieht man, daß dieses allgemeinere Konzept keinen Einfluß auf den Erwartungswert der Schätzfunktion hat. Bei gleicher, durch σ_1 gemessener Streuung der Werte ist im stochastischen

Modell $E(\hat{\theta}_1 - \theta_1)^2$ um $\text{var}(T_1)E(\beta(K)\frac{K}{n})^2$ größer; d.h., wenn man mit zufälligen Meßfehlern rechnen muß oder aus theoretischen Erwägungen die Werte als Realisierung von Zufallsvariablen ansieht, vergrößert sich $E(\hat{\theta}_1 - \theta_1)^2$ um den obengenannten Betrag gegenüber der Modellvorstellung, daß die Werte fest sind. Nur falls $\text{var}(T_1)=0$ ist, haben die beiden verschiedenen Modellvorstellungen keinen Einfluß auf den erwarteten Fehler $E(\hat{\theta}_1 - \theta_1)^2$. Dieser Fall kann im Modell der Superpopulation auftreten, wenn ein fester, unbekannter Betrag geschätzt werden soll, der nach einem Zufallsmechanismus auf die Untersuchungseinheiten der Grundgesamtheit aufgeteilt ist.

1.4 Zulässigkeit, vollständige Klassen von Schätz- funktionen

Für lineare, symmetrische Schätzfunktionen bei symmetrischen, von den Werten unabhängigen Stichprobenverfahren wurden in Satz 1.3.3 und 1.3.5 die Erwartungswerte der quadrierten Abweichung der Schätzfunktion vom wahren Wert, also die Werte der Risikofunktion bei Gaußscher Verlustfunktion für festes, beliebiges W bestimmt. In diesem Abschnitt sollen diese Ergebnisse genutzt werden, um Aussagen über zulässige Schätzfunktionen zu erhalten und um kleine vollständige Klassen von Schätzfunktionen zu bestimmen. Anschließend werden wir besonders auf θ_1^* und die Horvitz-Thompson Schätzfunktion $\tilde{\theta}_1$ eingehen und am Ende dieses Abschnittes Literatur zu dem Problem diskutieren, wobei wir auch bekannte Ergebnisse zur optimalen Wahl einer Schätzstrategie referieren werden.

Betrachtet wird nur die Klasse Ψ_1 der als Funktion der Zufallsvariablen $X_i S_i$ linearen symmetrischen Schätzfunktionen (vgl. Def. 1.2.3). Da die symmetrischen Stichprobenverfahren durch die Wahrscheinlichkeiten $\alpha_k = P(K=k), k=0, \dots, n$ bestimmt sind, hängt die Verteilung von linearen, symmetrischen Schätzfunktionen nur von den $\beta_k = \beta(k)$ mit $\alpha_k > 0$ ab; Aussagen über β_k beziehen sich deshalb immer nur auf die k mit $\alpha_k > 0$ und die Schätzfunktionen, die sich nur durch β_k mit $\alpha_k = 0$ unterscheiden, betrachten wir als eine Schätzfunktion. Mit β bezeichnen wir den $l \leq n$ dimensionalen Vektor, der aus den Komponenten β_k mit $\alpha_k > 0$ besteht. Wir werden nur Stichprobenverfahren betrachten, bei denen sicher eine nicht leere Stichprobe ausgewählt wird, d.h. wir werden $\alpha_0 = P(K=0) = 0$ setzen.

Nach Satz 1.3.5 hängt die Risikofunktion (bei Gaußscher Verlustfunktion) von der Verteilung W der Werte nur über die Parameter σ_1^2 , $\text{var}(T_1)$ und θ_1^2 ab; eine Familie $\mathcal{W}$ von Verteilungen können wir deshalb durch eine Parametermenge $\Lambda_1 \subset \mathbb{R}_+^3 := \{x \in \mathbb{R} : x \geq 0\}^3$ mit den Elementen $(\sigma_1^2, \text{var}(T_1), \theta_1^2)$

beschreiben. Bei den folgenden Beweisen muß der Parameter-
raum mindestens ein Element $(\sigma_1^2, \text{var}(T_1), \theta_1^2)$ mit $\sigma_1^2 \neq 0$ und $\theta_1^2 \neq 0$
enthalten, was allerdings keine wesentliche Einschränkung
bedeutet. Falls $\theta_1^2 = 0$ für alle Elemente von Λ_1 gilt, hat man
kein Schätzproblem mehr, sondern kennt den wahren Parameter;
falls $\sigma_1^2 = 0$ für alle Elemente von Λ_1 ist, gilt $X_i = \frac{T_1}{n}$ für
 $i=1,2,\dots,n$ und damit reduziert sich das Problem auf die
Schätzung von $E(T_1)$ aus einer Beobachtung von T_1 .

Die hier genannten Voraussetzungen werden im folgenden nicht
mehr explizit genannt, weiterhin setzen wir wie in Abschnitt
1.3 die Unabhängigkeit der Werte vom Stichprobenverfahren
voraus und legen die einfache Gaußsche Verlustfunktion
(Gl. (1.2.3)) zugrunde, falls nicht explizit von der logarith-
mischen Verlustfunktion ausgegangen wird. Nach Bemerkung
1.2.7 gelten die Ergebnisse auch für die allgemeine Gaußsche
Verlustfunktion (Gl. (1.2.4)) und damit speziell für die in
Gl. (1.2.5) definierte Verlustfunktion. Die Risikofunktion
läßt sich nach Satz 1.3.5 wie folgt darstellen:

$$(1.4.1) \quad R(\hat{\theta}_1, W) = R(\beta, \sigma_1^2, \text{var}(T_1), \theta_1^2) \\ = \sigma_1^2 g_1(\beta) + \text{var}(T_1) g_2(\beta) + \theta_1^2 g_3(\beta)$$

mit:

$$(1.4.2) \quad g_1(\beta) := E\left(\beta^2 (K) \frac{K(n-K)}{n-1}\right) = \sum_k \alpha_k \beta_k^2 \frac{k(n-k)}{n-1}$$

$$(1.4.3) \quad g_2(\beta) := E\left(\beta (K) \frac{K}{n}\right)^2 = \sum_k \alpha_k \beta_k^2 \frac{k^2}{n^2}$$

$$(1.4.4) \quad g_3(\beta) := E\left(\beta (K) \frac{K}{n} - 1\right)^2 = \sum_k \alpha_k \left(\beta_k^2 \frac{k^2}{n^2} - 2\beta_k \frac{k}{n} + 1\right),$$

wobei jeweils über alle k mit $\alpha_k > 0$ summiert wird.

Hilfssatz 1.4.1

Die Funktionen g_1 , g_2 und g_3 sind konvexe Funktionen von $\mathbb{R}^k$ in $\mathbb{R}$. Damit sind auch Linearkombinationen $c_1 g_1 + c_2 g_2 + c_3 g_3 + c_4$ mit $c_1, c_2, c_3, c_4 \in \mathbb{R}$; $c_1, c_2, c_3 \geq 0$ konvex und insbesondere ist die Risikofunktion R für feste $\sigma_1^2, \text{var}(T_1)$ und θ_1^2 konvex.

Beweis

Es ist zu zeigen, daß die Matrix der 2. partiellen Ableitungen positiv semidefinit ist (vgl. Luenberger (1973) S. 118). Es gilt:

$$(1.4.5) \quad \frac{\partial g_1}{\partial \beta_k} = 2\alpha_k \beta_k \frac{k(n-k)}{n-1}$$

$$(1.4.6) \quad \frac{\partial g_2}{\partial \beta_k} = 2\alpha_k \beta_k \frac{k^2}{n}$$

$$(1.4.7) \quad \frac{\partial g_3}{\partial \beta_k} = 2\alpha_k \beta_k \frac{k^2}{n^2} - 2\alpha_k \frac{k}{n}$$

$$(1.4.8) \quad \frac{\partial^2 g_i}{\partial \beta_k \partial \beta_j} = 0 \quad \text{für } i=1,2,3; k \neq j$$

$$(1.4.9) \quad \frac{\partial^2 g_1}{\partial \beta_k \partial \beta_k} = 2\alpha_k \frac{k(n-k)}{n-1} \geq 0$$

$$(1.4.10) \quad \frac{\partial^2 g_2}{\partial \beta_k \partial \beta_k} = 2\alpha_k \frac{k^2}{n^2} > 0$$

$$(1.4.11) \quad \frac{\partial^2 g_3}{\partial \beta_k \partial \beta_k} = 2\alpha_k \frac{k^2}{n^2} > 0$$

Für alle 3 Funktionen hat die Matrix der 2. partiellen Ableitungen Elemente ungleich 0 nur in der Diagonalen und diese Elemente sind größer gleich 0, also ist die Matrix positiv semidefinit und die g_i sind konvex. Die Aussage über die Konvexheit der Linearkombinationen folgt nach Luenberger (1973) S. 116.

Satz 1.4.2

(i) Durch

$$(1.4.12) \quad \beta_k = \frac{1}{(n-k)a+kb}, \quad a \geq 0, \quad b \geq \frac{1}{n}, \quad k=1, \dots, n$$

definierte lineare Schätzfunktionen sind gleichmäßig beste Schätzfunktionen bzgl. dem Parameter-
raum $\Lambda_{a,b} := \{(\theta_1^2 a \frac{n-1}{n}, \theta_1^2 (bn-1), \theta_1^2) : \theta_1^2 > 0\}$.
Falls $\Lambda_1 \cap \Lambda_{a,b} \neq \emptyset$, sind sie zulässig bzgl. Λ_1

(ii) Falls Λ_1 Parameter $(\sigma_1^2, \text{var}(T_1), 0)$ mit $\text{var}(T_1) > 0$ enthält, ist die durch $\beta_k = 0, k=1, \dots, n$ definierte Schätzfunktion bzgl. Λ_1 zulässig.

(iii) Falls Λ_1 Parameter $(\sigma_1^2, 0, 0)$ mit $\sigma_1^2 > 0$ enthält, sind die durch $\beta_k = 0, k=1, \dots, n-1$ und $\beta_n = \frac{\theta_1^2}{\text{var}(T_1) + \theta_1^2}$ mit $(\sigma_1^2, \text{var}(T_1), \theta_1^2) \in \Lambda_1$ und $\theta_1^2 > 0$ definierten Schätzfunktionen zulässig.

Beweis

Für feste Parameter $(\sigma_1^2, \text{var}(T_1), \theta_1^2)$ mit $\theta_1^2 > 0$ oder $\text{var}(T_1) > 0$ gilt:

$$(1.4.13) \quad \frac{\partial R}{\partial \beta_k} = 2\alpha_k \left(\sigma_1^2 \beta_k \frac{k(n-k)}{n-1} + \text{var}(T_1) \beta_k \frac{k^2}{n^2} + \theta_1^2 \beta_k \frac{k^2}{n^2} - \theta_1^2 \frac{k}{n} \right)$$

$$(1.4.14) \quad \frac{\partial^2 R}{\partial \beta_k \partial \beta_j} = 0 \quad \text{für } k \neq j$$

$$(1.4.15) \quad \frac{\partial^2 R}{\partial \beta_k \partial \beta_k} = 2\alpha_k \left(\sigma_1^2 \frac{k(n-k)}{n-1} + \text{var}(T_1) \frac{k^2}{n^2} + \theta_1^2 \frac{k^2}{n^2} \right) > 0$$

Die einzige Nullstelle der 1. Ableitung lautet:

$$(1.4.16) \quad \beta_k = \frac{\theta_1^2}{\sigma_1^2 \frac{n(n-k)}{n-1} + (\text{var}(T_1) + \theta_1^2) \frac{k}{n}}, \quad k=1, \dots, n; \alpha_k > 0$$

Da die Matrix der 2. partiellen Ableitungen positiv definit ist und da R konvex ist, ist dies das einzige globale Minimum (vgl. Luenberger (1973) S. 114, 119).

zu (i): Für $\theta_1^2 > 0$ kann man (1.4.16) auch in der Form (1.4.12) mit $a = \frac{\sigma_1^2}{\theta_1^2} \frac{n}{n-1}$ und $b = \frac{\text{var}(T_1) + \theta_1^2}{n\theta_1^2}$ schreiben und die durch β_k definierte Schätzfunktion ist für $(\sigma_1^2, \text{var}(T_1), \theta_1^2) \in \Lambda_{a,b}$ die einzige gleichmäßig beste Schätzfunktion und daher nach Bemerkung 1.2.6 bzgl. Λ_1 zulässig, falls $\Lambda_{a,b} \cap \Lambda_1 \neq \emptyset$.

zu (ii): Falls $\theta_1^2 = 0$ und $\text{var}(T_1) > 0$, gilt in Gl. (1.4.16) $\beta_k = 0$ für $k=1, \dots, n$ und nach Bemerkung 1.2.6 ist die so definierte Schätzfunktion zulässig.

zu (iii): Es sei $\alpha_n = P(K=n) > 0$. (Falls $\alpha_n = 0$, kann β beliebig definiert werden, ohne die Verteilung und damit die Güteeigenschaften der Schätzfunktion zu verändern). β sei gemäß Teil (iii) des Satzes definiert und

$\tilde{\beta} \in \mathbb{R}^l$ sei ein beliebiger Vektor. Wir zeigen, daß die durch $\tilde{\beta}$ definierte Schätzfunktion nicht gleichmäßig besser als die durch β definierte Schätzfunktion sein kann. Für $\sigma_1^2 > 0$ gilt $0 = R(\beta, \sigma_1^2, 0, 0) \geq R(\tilde{\beta}, \sigma_1^2, 0, 0)$ genau dann, wenn $\tilde{\beta}_k = 0$ für $k=1, \dots, (n-1)$ mit $\alpha_k > 0$. Für $\tilde{\beta}$, die dieser Bedingung genügen, gilt:

$$(1.4.17) \quad R(\tilde{\beta}, \sigma_1^2, \text{var}(T_1), \theta_1^2) = (\text{var}(T_1) \tilde{\beta}_n^2 + \theta_1^2 (\tilde{\beta}_n - 1)^2) \alpha_n$$

Für feste Parameter $(\sigma_1^2, \text{var}(T_1), \theta_1^2)$ mit $\theta_1^2 > 0$ ist dies eine konvexe Funktion von $\tilde{\beta}_n$; durch Differenzieren erhält man als einziges globales Minimum:

$$\beta_n = \frac{\theta_1^2}{\text{var}(T_1) + \theta_1^2}. \text{ Alle Schätzfunktionen, die nicht den}$$

Bedingungen des Teils (iii) des Satzes genügen, sind also für Parameter $(\sigma_1^2, 0, 0)$ oder $(\sigma_1^2, \text{var}(T_1), \theta_1^2)$ schlechter als die durch $\beta_k = 0$ für $k=1, \dots, (n-1)$ und

$$\beta_n = \frac{\theta_1^2}{\text{var}(T_1) + \theta_1^2} \text{ definierte Schätzfunktion.}$$

Bemerkung 1.4.3

Die Festlegung der β_k in Satz 1.4.2 ist nur für $\alpha_k > 0$ wesentlich, für k mit $\alpha_k = 0$ können sie beliebig anders gewählt werden. Deswegen ist in (iii) die Festlegung von β_n nur wesentlich, wenn mit positiver Wahrscheinlichkeit alle Elemente der Grundgesamtheit ausgewählt werden können; in diesem Fall ist auch die mögliche Festlegung $\beta_n = \frac{\theta_1^2}{0 + \theta_1^2} = 1$ intuitiv klar: für $S = \Omega$ und $\text{var}(T_1) = 0$ ist die Summe der beobachteten Werte ($\beta_n = 1$) gleich dem wahren Wert θ_1 .

Bemerkung 1.4.4

Falls bei festen Werten ($\text{var}(X_i)=0$ stets) der Variationskoeffizient $v_1 = \frac{\sigma_1}{\theta_1}$ bekannt ist oder allgemeiner $v_1 = \frac{\sigma_1}{\theta_1}$ und $w_1^2 = \frac{\text{var}(T_1)}{\theta_1^2}$ bekannt sind und $\theta_1^2 > 0$ gilt, existieren gleichmäßig beste Schätzfunktionen, die durch

$$(1.4.18) \quad \beta_k = \frac{1}{v_1^2 \frac{n(n-k)}{n-1} + (w_1^2 + 1) \frac{k}{n}} \quad \text{definiert sind.}$$

Satz 1.4.5

Die Menge der linearen, symmetrischen Schätzfunktionen mit

$$(1.4.19) \quad 0 \leq \beta_k \leq \frac{n}{k} \quad \text{für } k = 1, \dots, n$$

ist vollständig.

Beweis

Gegeben sei eine Schätzfunktion durch $\tilde{\beta}$ mit $\tilde{\beta}_k \in [0, \frac{n}{k}]$ für mindestens ein k mit $\alpha_k > 0$. Falls $-\frac{n}{k} \leq \tilde{\beta}_k < 0$, setze man $\beta_k = |\tilde{\beta}_k|$, falls $|\tilde{\beta}_k| > \frac{n}{k}$ setze man $\beta_k = \frac{n}{k}$ und sonst $\tilde{\beta}_k = \beta_k$. Wegen $g_1(\tilde{\beta}) \geq g_1(\beta)$, $g_2(\tilde{\beta}) \geq g_2(\beta)$ und $g_3(\tilde{\beta}) > g_3(\beta)$ und gilt für die Risikofunktion:

$$(1.4.20) \quad R(\tilde{\beta}, \sigma_1^2, \text{var}(T_1), \theta_1^2) \geq R(\beta, \sigma_1^2, \text{var}(T_1), \theta_1^2)$$

für alle $(\sigma_1^2, \text{var}(T_1), \theta_1^2) \in \mathbb{R}_+^3$ mit Ungleichheit bei $\theta_1^2 \neq 0$.

Bemerkung 1.4.6

Falls die Stichprobengröße nicht fest vorgegeben ist, d.h. $\text{var}(K) > 0$, gehört die Horvitz-Thompson-Schätzfunktion nicht zu der in Satz 1.4.5 angegebenen vollständigen Menge, weil $\beta_k = \frac{n}{E(K)} > \frac{n}{k}$ für $k > E(K)$ gilt. Es gibt also eine lineare, symmetrische Schätzfunktion, die gleichmäßig besser als die Horvitz-Thompson-Schätzfunktion ist.

Satz 1.4.7

- (i) Die Menge der in Satz 1.4.2 definierten linearen, symmetrischen Schätzfunktionen, d.h.:

$$(1.4.21) \quad \Psi_1' := \{ \hat{\theta}_1 = \beta(K) \sum_{i=1}^n X_i S_i : \beta(K) = \frac{1}{(n-K)a + Kb} \text{ mit } a \geq 0$$

$$\text{und } b \geq \frac{1}{n}$$

$$\text{oder } \beta_k = \beta(k) = 0 \text{ für } k \leq n-1 \text{ und } \beta_n \in [0, 1] \}$$

ist vollständig.

- (ii) Falls $\Lambda_1 = \mathbb{R}_+^3$, ist Ψ_1' die Menge aller zulässigen Schätzfunktionen und damit die minimale vollständige Klasse.

- (iii) Für $\Lambda_1 \subset \{ (\sigma_1^2, \text{var}(T_1), \theta_1^2) : \sigma_1^2 \geq 0, \theta_1^2 \geq 0, \text{var}(T_1) = 0 \}$ (d.h. für feste, nicht zufällige Werte) ist die Klasse

$$(1.4.22) \quad \Psi_1'' := \{ \hat{\theta}_1 = \beta(K) \sum_{i=1}^n X_i S_i : \beta(K) = \frac{1}{(n-K)a + \frac{K}{n}} \text{ mit } a \geq 0$$

$$\text{oder } \beta_k = \beta(k) = 0 \text{ für } k \leq n-1 \text{ und } \beta_n = 1 \}$$

vollständig; für $\Lambda_1 = \{(\sigma^2, \text{var}(T), \theta^2) : \sigma^2 \geq 0, \theta^2 \geq 0, \text{var}(T) = 0\}$ ist Ψ_1'' die Menge aller zulässigen Schätzfunktionen und damit die minimale vollständige Klasse.

Beweis

zu (i): Bei der Betrachtung können wir uns auf die in Satz 1.4.5 angegebene vollständige Klasse von Schätzfunktionen beschränken, d.h. ohne Einschränkung der Allgemeinheit Gl. (1.4.19) voraussetzen. Für eine durch $\tilde{\beta}$ definierte Schätzfunktion nicht aus Ψ_1'' gilt $g_1(\tilde{\beta}) > 0$ und $g_2(\tilde{\beta}) > 0$. Wir konstruieren eine bessere Schätzfunktion, indem wir $g_3(\beta)$ unter Nebenbedingungen

$$(1.4.23) \quad g_1(\beta) - g_1(\tilde{\beta}) \leq 0 \quad \text{und}$$

$$(1.4.24) \quad g_2(\beta) - g_2(\tilde{\beta}) \leq 0 \quad \text{minimieren.}$$

Die durch die Gl. (1.4.19), (1.4.23) und (1.4.24) definierte Teilmenge von $\mathbb{R}^k$ ist kompakt, daher existiert ein Minimum von $g_3(\beta)$. Außerdem sind die Funktionen $g_3(\beta)$, $g_1(\beta) - g_1(\tilde{\beta})$, $g_2(\beta) - g_2(\tilde{\beta})$ konvex (vgl. Hilfssatz 1.4.1), die durch Gl. (1.4.19) definierte Teilmenge des $\mathbb{R}^k$ ist konvex und wegen $g_1(\tilde{\beta}) > 0$ und $g_2(\tilde{\beta}) > 0$ existiert ein β mit $g_1(\beta) - g_1(\tilde{\beta}) < 0$ und $g_2(\beta) - g_2(\tilde{\beta}) < 0$ (z.B. $\beta = \frac{1}{2}\tilde{\beta}$). Nach dem Satz von Kuhn-Tucker (vgl. Ioffe, Tichomirov (1979) S. 72) ist jede Lösung des Minimierungsproblems ein Minimum der Lagrangefunktion

$$(1.4.25) \quad L(\beta, \lambda_1, \lambda_2) := g_3(\beta) + \lambda_1(g_1(\beta) - g_1(\tilde{\beta})) + \lambda_2(g_2(\beta) - g_2(\tilde{\beta}))$$

für bestimmte "Lagrange-Multiplikatoren" $\lambda_1 \geq 0, \lambda_2 \geq 0$.

Beim Vergleich der Lagrangefunktion als Funktion von β für feste λ_1, λ_2 mit der Risikofunktion sehen wir

$$(1.4.26) \quad L(\beta, \lambda_1, \lambda_2) = R(\beta, \lambda_1, \lambda_2, 1) - \lambda_1 g_1(\tilde{\beta}) - \lambda_2 g_2(\tilde{\beta});$$

die Lagrangefunktion unterscheidet sich also nur um eine Konstante von der Risikofunktion für bestimmte Parameter.

Wie im Beweis von Satz 1.4.2 erhält man als einziges Minimum der Lagrangefunktion:

$$(1.4.27) \quad \beta_k = \frac{1}{(n-k)a+kb} \quad \text{mit} \quad a = \frac{n}{n-1}\lambda_1 \geq 0 \quad \text{und} \quad b = \frac{\lambda_2+1}{n} \geq \frac{1}{n}$$

Für diese β_k gilt stets Gl. (1.4.19). Alle Minima β von $g_3(\beta)$ unter den Nebenbedingungen (1.4.23) und (1.4.24) müssen also die Form (1.4.27) haben und weil die durch $\tilde{\beta}$ definierte Schätzfunktion nicht aus Ψ'_1 ist, gilt:

$$(1.4.28) \quad \begin{aligned} R(\beta, \sigma_1^2, \text{var}(T_1), \theta_1^2) - R(\tilde{\beta}, \sigma_1^2, \text{var}(T_1), \theta_1^2) \\ = \sigma_1^2 (g_1(\beta) - g_1(\tilde{\beta})) + \text{var}(T_1) (g_2(\beta) - g_2(\tilde{\beta})) \\ + \theta_1^2 (g_3(\beta) - g_3(\tilde{\beta})) \leq 0 \end{aligned}$$

mit Ungleichheit für $\theta_1^2 \neq 0$.

zu (ii): Dieser Teil folgt direkt aus (i), aus Satz 1.4.2 und Bemerkung 1.2.5.

zu (iii): Es sei Ψ_1''' die Menge aller Schätzfunktionen mit $\beta_k=0$ für $k=1, \dots, (n-1)$. Nach dem Beweis von Satz 1.4.2 (iii) und wegen $\text{var}(T)=0$ ist die durch $\beta_k=0$ für $k \leq n-1$ und $\beta_n=1$ definierte Schätzfunktion die

gleichmäßig beste in Ψ_1''' . Für Schätzfunktionen $\tilde{\theta} \in \Psi_1'' \cup \Psi_1'''$ läßt sich analog zum Beweis von Teil (i) dieses Satzes eine bessere Schätzfunktion aus Ψ_1'' konstruieren. Allerdings fällt in diesem Fall die Nebenbedingung (1.4.24) und damit der zweite Lagrange-Multiplikator λ_2 weg (d.h. $\lambda_2 = 0$ und somit $b = \frac{1}{n}$). Die Aussage zur Minimalität von Ψ_1'' folgt wieder aus Satz 1.4.2 und Bemerkung 1.2.5.

Bemerkung 1.4.8

Bei der Wahl einer Schätzfunktion kann man sich also auf Schätzfunktionen aus Ψ_1' bzw. Ψ_1'' beschränken. Für Parameter-räume $\Lambda_1 \neq \mathbb{R}_+^3$, $\Lambda_1 \neq \Lambda_{a,b}$ oder $\Lambda_1 \neq \{(\sigma^2, \text{var}(T), \theta^2) : \sigma^2 \geq 0, \theta^2 \geq 0, \text{var}(T) = 0\}$ haben wir weder in Satz 1.4.2 alle zulässigen Schätzfunktionen noch in Satz 1.4.7 die minimale vollständige Klasse bestimmt. Dieses Problem könnte man möglicherweise mit weiteren Bedingungen an a und b (Gl. (1.4.21)) lösen.

Allerdings sind die Klassen Ψ_1' und Ψ_1'' schon relativ einfach und man kann mit der Wahl von a und b in Gl. (1.4.27) die Gewichtungen $g_1(\beta)$ von σ_1^2 , $g_2(\beta)$ von $\text{var}(T_1)$ und $g_3(\beta)$ von θ_1^2 so bestimmen, daß bestimmte Parameter wenig Einfluß auf die Werte der Risikofunktion erhalten. Die zulässigen Schätzfunktionen mit $\beta_k \neq 0$ für $k \leq n-1$ wird man kaum auswählen, da sie für die wichtigsten Stichproben $S \neq \Omega$ nicht von den beobachteten Werten abhängen; der Grund für ihre Zulässigkeit liegt im wesentlichen darin, daß ein Parameter $(\sigma_1^2, \text{var}(T_1), 0)$ mit $\sigma_1^2 > 0$ in der abgeschlossenen Hülle des Parameterraums Λ_1 liegt (vgl. Beweis von Satz 1.4.2 (ii) und (iii)).

Bemerkung 1.4.9

Es gibt nur eine erwartungstreue Schätzfunktion in Ψ_1' bzw. $\Psi_1'' : \theta_1^* \text{ mit } \beta^*(K) = \frac{n}{K}$. Für alle anderen Schätzfunktionen $\hat{\theta}_1 \in \Psi_1'$ gilt $0 \leq \beta(K) < \beta^*(K)$ und somit

$$(1.4.29) \quad |E(\hat{\theta}_1)| < |E(\theta_1^*)| = |\theta_1|$$

Die Schätzfunktionen verzerren also nach 0 hin.

Bemerkung 1.4.10

Außer der in der vorigen Bemerkung erwähnten Erwartungstreue hat θ_1^* noch mehrere andere Vorteile: der Wert der Risikofunktion hängt nicht vom wahren Wert θ_1 ab; unter der Bedingung, daß eine Stichprobe vom Umfang k gezogen wurde, ist θ_1^* immer noch erwartungstreu, d.h.:

$$(1.4.30) \quad E(\theta_1^* | K) = \theta_1;$$

und nicht zuletzt hat θ_1^* eine einfache Form und ist als n -faches des arithmetischen Mittels der beobachteten Werte gut zu interpretieren. Außer θ_1^* haben wir bisher nur eine andere erwartungstreue Schätzfunktion behandelt, die Horvitz-Thompson Schätzfunktion $\tilde{\theta}_1$ mit $\beta_k = \frac{n}{E(K)}$. Gegenüber θ_1^* hat $\tilde{\theta}_1$ den Vorteil, daß σ_1^2 geringer durch $g_1(\beta)$ gewichtet wird, wie man mit Hilfe der Jensenschen Ungleichung (vgl. Witting (1974) S. 208) zeigen kann. Falls also σ_1^2 im Verhältnis zu θ_1^2 und $\text{var}(T_1)$ genügend groß ist, ist $\tilde{\theta}_1$ besser als θ_1^* . Allerdings gehört $\tilde{\theta}_1$ nicht zu Ψ_1' ; wenn man also bereit ist, auf die Erwartungstreue zu verzichten, kann man eine gleichmäßig bessere Schätzfunktion als $\tilde{\theta}_1$ in Ψ_1' finden. Außerdem gilt die Erwartungstreue bei der Horvitz-Thompson Schätzfunktion in einem relativ schwachen Sinn, nach Ziehung der Stichprobe, d.h. nach der Realisierung der Zufallsvariablen K muß man mit einer systematischen Verzerrung rechnen; genauer gesagt gilt:

$$(1.4.31) \quad E(\tilde{\theta}_1 | K=k) \neq \theta_1 \quad \text{für} \quad k \neq E(K) \quad \text{und} \quad \theta_1 \neq 0.$$

Zusammenfassend kann man sagen, daß es unter den gegebenen Voraussetzungen wenig Gründe für die Benutzung der Horvitz-Thompson Schätzfunktion gibt und daß unter den Schätzfunktionen aus Ψ_1' bzw. Ψ_1'' die naheliegenste und plausibelste Schätzfunktion θ_1^* ist.

In dem folgenden Satz zeigen wir, daß θ_1^* bzgl. der logarithmischen Verlustfunktion (vgl. Gl. (1.2.6)) gleichmäßig beste Schätzfunktion ist. Da in Abschnitt 1.2 nur eine relativ schwache inhaltliche Begründung für die Wahl der logarithmischen Verlustfunktion gegeben wurde, liegt der Sinn des Satzes weniger in der Angabe einer gleichmäßig besten Schätzfunktion als umgekehrt in einer Beschreibung der Bewertung der Schätzfehler, die man impliziert, wenn man θ_1^* allen anderen Schätzfunktionen vorzieht.

Satz 1.4.11

Bei Zugrundelegung der logarithmischen Verlustfunktion (Gl. (1.2.6)) ist θ_1^* mit $\beta_k^* = \frac{n}{k}, k=1, \dots, n$ gleichmäßig beste Schätzfunktion, wobei hier $\theta_1 > 0$ und für die Schätzfunktion $P(\hat{\theta}_1 > 0) = 1$ vorausgesetzt werden muß.

Beweis

Für die logarithmische Verlustfunktion mit $g(\theta) \equiv 1$ gilt nach Gl. (1.3.3)

$$\begin{aligned} (1.4.32) \quad R(\beta, \sigma_1^2, \text{var}(T_1), \theta^2) &= E\left(\frac{\hat{\theta}_1}{\theta_1}\right) - E\left(\log \frac{\hat{\theta}_1}{\theta_1}\right) - 1 \\ &= E\left(\beta(K) \frac{K}{n}\right) - E\left(\log \beta(K) \frac{K}{n}\right) - 1 \geq 0 \end{aligned}$$

Da aus $P(\hat{\theta}_1 > 0) = 1$ $\beta_k > 0$ für alle k mit $\alpha_k > 0$ folgt, ist die Risikofunktion definiert und sie ist unabhängig von $\sigma_1^2, \text{var}(T_1)$ und θ_1^2 . Ihr Minimum 0 nimmt sie nur für $\beta^*(K) = \frac{n}{K}$ an und daher gilt:

$$(1.4.33) \quad 0 = R(\beta^*, \sigma_1^2, \text{var}(T_1), \theta_1^2) \leq R(\beta, \sigma_1^2, \text{var}(T_1), \theta_1^2) \\ \text{für alle } \beta_k > 0, \sigma_1^2 \geq 0, \text{var}(T_1) \geq 0, \theta_1^2 > 0.$$

Bemerkung 1.4.12

In "Loss functions and admissible estimators in survey sampling" (1979) bestimmt Stenger die Menge aller zulässigen linearen symmetrischen Schätzfunktionen für das arithmetische Mittel $\mu_1 = \frac{1}{n} \theta_1$ der Werte bei symmetrischen Stichproben. Er geht von festen, positiven Werten $X_i \equiv x_i$ aus, d.h. es gilt

$\text{var}(X_i) = 0$, $\text{var}(T_1) = 0$ und $\sigma_1^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu_1)^2$. Als mögliche Werte der x_i läßt er $\mathbb{R}_+$, $\mathbb{R}_+ \setminus \{0\}$, $\{0, 1\}$, $\{0, 1, 2, \dots\}$ oder $\{1, 2, \dots\}$ zu. Unter diesen Voraussetzungen gilt für den Parameterraum:

$$(1.4.34) \quad \Lambda_1 = \{(\sigma_1^2, 0, \theta_1^2) : 0 \leq \sigma_1^2 \leq \frac{n-1}{2} \theta_1^2, \theta_1^2 \geq 0 \text{ (bzw. } \theta_1^2 > 0)\}$$

und im Parameterraum ist insbesondere kein Element $(\sigma_1^2, 0, 0)$ mit $\sigma_1^2 > 0$ enthalten. Nach Satz 1.4.2 bestimmt man zulässige Schätzfunktionen für θ_1 durch

$$(1.4.35) \quad \beta_k = \frac{1}{(n-k)a + \frac{k}{n}} \quad \text{mit} \quad 0 \leq a \leq \frac{1}{n}$$

Wegen Bemerkung 1.2.8 haben die entsprechenden zulässigen Schätzfunktionen $\hat{\mu}_1$ für $\mu_1 = \frac{1}{n} \theta$ die Form:

$$(1.4.36) \quad \hat{\mu}_1 = \frac{1}{n(n-K)a+K} \sum_{i=1}^n x_i S_i \quad \text{mit} \quad 0 \leq a \leq \frac{1}{n}.$$

Stenger beweist nun zusätzlich, daß durch Gl. (1.4.36) die Menge aller zulässigen Schätzfunktionen definiert wird, wobei er allerdings den Koeffizienten $\frac{1}{n(n-k)a+k}$ in einer anderen Form schreibt. Zur logarithmischen Verlustfunktion beweist Stenger für positive, feste Werte $x_i > 0$ eine Umkehrung des Satzes 1.4.11; wenn θ_1^* die einzige zulässige Schätzfunktion in der Klasse der linearen symmetrischen Schätzfunktion ist bezüglich einer Verlustfunktion L mit stetiger partieller Ableitung nach dem ersten Argument $\hat{\theta}_0$, dann ist L die logarithmische Verlustfunktion.

Bemerkung 1.4.13

Wir haben bisher das Stichprobenverfahren als fest vorgegeben vorausgesetzt; in dieser Bemerkung sollen Ergebnisse von Stenger (1981) und (1975) zur optimalen Wahl eines symmetrischen Stichprobenverfahrens P bzw. einer Schätzstrategie referiert werden. Beide Arbeiten beziehen sich auf feste Werte $X_i \equiv x_i, i=1, \dots, n$ und geschätzt werden soll das Mittel der Werte $\mu_1 = \frac{1}{n} \theta_1 = \frac{1}{n} \sum_{i=1}^n x_i$. In der Arbeit "On the completeness of the class of fixed size sampling strategies" betrachtet Stenger ausschließlich die Schätzfunktion $\mu_1^* = \frac{1}{n} \theta_1^*$ und bestimmt eine vollständige Klasse von Stichprobenverfahren, wobei die Vollständigkeit von Klassen von Stichprobenverfahren analog zu der von Schätzfunktionen in Def. 1.2.4 definiert wird. Die Risikofunktion habe die Form

$$(1.4.37) \quad R(P, \mu_1) = E(L_1(\mu_1^*, x_1, \dots, x_n) + cK)$$

mit $c > 0$. Unter c kann man sich die konstanten Kosten

für die Ziehung eines Elementes vorstellen. Die Funktion L von $\mathbb{R}^{n+1}$ in $\mathbb{R}$ sei als Funktion des ersten Argumentes in einem Intervall $I \subset \mathbb{R}$ für alle $x = (x_1, \dots, x_n)$ streng konvex und die zweite Ableitung existiere und sei konvex. Außerdem gelte $x_i \in I$ für $i=1, \dots, n$. Unter diesen Bedingungen ist die Menge der Stichprobenverfahren der Form

$$(1.4.38) \quad P^* = \alpha_k P^{(k)} + (1 - \alpha_k) P^{(k+1)}, \quad k \in \{1, \dots, (n-1)\} \quad 0 \leq \alpha_k \leq 1$$

vollständig, wobei $P^{(k)}$ das k -malige Ziehen ohne Zurücklegen bedeutet (siehe Def. 1.1.4). Zu jedem Stichprobenverfahren, bei dem der Umfang K mehr als 2 verschiedene Werte annehmen kann, gibt es ein gleichmäßig besseres Verfahren, bei dem K höchstens 2 benachbarte Werte annimmt.

In der Arbeit von 1975 betrachtet Stenger Schätzstrategien $(\hat{\mu}_1, P)$, wobei $\hat{\mu}_1$ lineare symmetrische Schätzfunktionen für das arithmetische Mittel μ_1 und P symmetrische Stichprobenverfahren mit $E(K)=e$ konstant sind. Wenn die x_i Werte größer gleich 0 annehmen können, erhält er als minimale vollständige Klasse von Schätzstrategien alle $(\hat{\mu}_1, P^*)$, für die $\hat{\mu}_1$ Gl.

(1.4.36) und P^* Gl. (1.4.38) mit $E(K)=e$ genügt. Falls man sich das Stichprobenverfahren auswählen kann, wird man also das Ziehen ohne Zurücklegen oder ein Verfahren wählen, bei dem der Stichprobenumfang K höchstens zwei benachbarte Werte annehmen kann. Es sei allerdings erwähnt, daß unter anderen Voraussetzungen, z.B. wenn die x_i alle Werte aus $\mathbb{R}$ annehmen können, auch andere Stichprobenverfahren zulässig sein können (vgl. Stenger (1972b)).

Bemerkung 1.4.14

Die wichtigste Rolle unter den symmetrischen Stichprobenverfahren spielt das k -malige Ziehen ohne Zurücklegen, das

nach Satz 1.1.8 sozusagen der Grundbaustein aller anderen symmetrischen Verfahren ist. Auch wenn nicht wie in der vorigen Bemerkung Optimalitätseigenschaften dieses Verfahrens bekannt sind, wird es vom Anwender oft gewählt, weil es analytisch recht einfach ist ($\alpha_k=1$, $E(K)=k$ usw.) oder weil der Anwender die Stichprobengröße im voraus festlegen will. Die in Satz 1.4.7 angegebenen vollständigen Klassen von Schätzfunktionen lauten in diesem Fall:

$$(1.4.39) \quad \psi_1' = \psi_1'' = \{\hat{\theta}_1 : 0 \leq \beta_k \leq \frac{n}{k}\} ;$$

Satz 1.4.7 bringt also gegenüber Satz 1.4.5 keinen wesentlichen Informationsgewinn. Außerdem gilt dann für die Horvitz-Thompson Schätzfunktion $\tilde{\theta}_1 = \theta_1^*$ und der Vergleich von θ_1^* und $\tilde{\theta}_1$ in Bemerkung 1.4.10 wird damit gegenstandslos.

Wir wollen noch den Fall betrachten, daß das Stichprobenverfahren nur 2 Stichprobenumfänge k_1 und k_2 realisieren kann, d.h. $\alpha_{k_1}, \alpha_{k_2} > 0$ und $\alpha_{k_1} + \alpha_{k_2} = 1$. Für $\beta_{k_1} \neq 0$ und $\beta_{k_2} \neq 0$ kann man zwar die Gleichungen

$$(1.4.40) \quad \beta_k = \frac{1}{(n-k)a+kb} ; \quad a, b \in \mathbb{R}; \quad k=k_1, k_2$$

nach a und b auflösen, für die Lösung (a_0, b_0) kann allerdings $a_0 < 0$ oder $b_0 < \frac{1}{n}$ gelten, auch wenn $0 < \beta_k \leq \frac{n}{k}$ ist. In diesem Fall unterscheidet sich ψ_1' von der in Satz 1.4.5 genannten vollständigen Klasse.

1.5 Die Schätzung anderer Parameter

In diesem Abschnitt sollen Möglichkeiten aufgezeigt werden, die in der Risikofunktion vorkommenden Parameter σ_1^2 , $\text{var}(T_1)$ und θ_1^2 erwartungstreu zu schätzen. Falls man nun Schätzwerte für diese Parameter hat, kann man sie verwenden, um für die bisher behandelten Schätzfunktionen $\hat{\theta}_1$ das Risiko bzw. bei erwartungstreuen Schätzfunktionen die Varianz abzuschätzen oder man hat die Möglichkeit, zulässige lineare symmetrische Schätzfunktionen gemäß Satz 1.4.2 so zu wählen, daß sie für die geschätzten Werte der Parameter σ_1^2 , $\text{var}(T_1)$, θ_1^2 optimal sind. Außerdem ist es oft für sich genommen interessant, die durch σ_1^2 oder σ_{11}^2 gemessene Streuung der Werte in der Grundgesamtheit oder bei zufälligen Werten die Varianz ihrer Summe $\text{var}(T_1)$ bzw. die Summen der Varianzen und Kovarianzen $\sum_{i=1}^n \text{var}(X_i)$ und $\sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \text{cov}(X_i, X_j)$ zu schätzen.

Um den Rahmen dieser Arbeit nicht zu übersteigen, wird zu jedem Parameter jeweils nur eine erwartungstreu Schätzfunktion angegeben und auf die Untersuchung weiterer Optimalitätseigenschaften dieser Schätzfunktionen verzichtet; es handelt sich also nur um einen ersten Schritt, Aussagen über diese Parameter aus der Stichprobe zu erhalten. Die hier entwickelten Schätzfunktionen sind von ähnlichem Typ wie θ_1^* ; es handelt sich um Schätzfunktionen, die auf Stichprobenmittelwerten basieren. Andere lineare Schätzfunktionen, wie z.B. die Horvitz-Thompson Schätzfunktionen, werden nicht behandelt. Bei der Schätzung der Parameter a_{11} und a_{12} werden wir allerdings noch darauf eingehen.

Wir beginnen mit der Schätzung der in Satz 1.3.3 definierten Parameter a_{11} und a_{12} und können aus diesen Schätzfunktionen

mittels Gl. (1.3.15) eine erwartungstreue Schätzfunktion für σ_1^2 herleiten. a_{11} ist der Erwartungswert der Summe der x_i^2 ; wir können also die bisherigen Ergebnisse verwenden und a_{11} mittels linearer, symmetrischer Schätzfunktionen schätzen. Dabei wählen wir speziell die θ_1^* entsprechende, erwartungstreue Schätzfunktion:

$$(1.5.1) \quad a_{11}^* = \frac{n}{K} \sum_{i=1}^n x_i^2 S_i.$$

Die Horvitz-Thompson Schätzfunktion lautet

$$(1.5.2) \quad \tilde{a}_{11} = \frac{n}{E(K)} \sum_{i=1}^n x_i^2 S_i;$$

da wir in Bemerkung 1.4.10 begründet haben, daß sie nicht sehr geeignet und unzulässig ist, werden wir sie nicht weiter verwenden.

Lineare, symmetrische Schätzfunktionen für a_{12} können wir allgemein mit Hilfe des nächsten Kapitels erhalten und diskutieren (Abschnitte 2.2 und 2.3). Da in diesem Fall wieder die Horvitz-Thompson Schätzfunktionen unzulässig sind, behandeln wir hier ausschließlich folgende Schätzfunktion:

$$(1.5.3) \quad a_{12}^* = \frac{n(n-1)}{K(K-1)} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n x_i x_j S_i S_j,$$

wobei wegen der Unabhängigkeit der Werte von Stichprobenverfahren gilt:

$$(1.5.4) \quad E(a_{12}^*) = \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n E(x_i x_j) E\left(\frac{S_i S_j}{K(K-1)}\right) n(n-1) = a_{12}$$

Weil σ_1^2 nach Gl. (1.3.15) eine Linearkombination von a_{11} und a_{12} ist, erhalten wir als erwartungstreue Schätzfunktion für σ_1^2 :

$$(1.5.5) \quad \sigma_1^{2*} = \frac{n-1}{n} a_{11}^* - \frac{1}{n} a_{12}^*$$

$$= \frac{n-1}{n} \frac{1}{K-1} \sum_{i=1}^n \left(X_i - \frac{1}{K} \sum_{j=1}^n X_j S_j \right)^2 S_i ,$$

also die mit $\frac{n-1}{n} \frac{1}{K-1}$ gewichtete Summe der quadrierten Abweichungen der beobachteten Werte vom Stichprobenmittel. Falls $X_i \equiv x_i$ feste Werte sind, ist dies die in der Stichprobentheorie übliche Schätzung der Streuung σ_1^2 beim Ziehen ohne Zurücklegen, hier wird diese Schätzfunktion für alle symmetrischen Stichprobenverfahren genommen. Beim "Ziehen mit Zurücklegen" wird in der Literatur meistens eine andere Schätzfunktion für σ_1^2 angegeben; beim Vergleich dieser Schätzfunktion mit der hier definierten ist darauf zu achten, daß im Gegensatz zur Notation dieser Arbeit in Lehrbüchern der Stichprobentheorie mit x_i oder y_i oft der Wert der Untersuchungseinheit bezeichnet wird, die im i -ten Zug gezogen wird. Im Unterschied zu dieser Arbeit können sich dann beim "Ziehen mit Zurücklegen" Werte x_i und x_j mit $i \neq j$ auf die gleiche Untersuchungseinheit beziehen, wenn im i -ten und j -ten Zug die gleiche Untersuchungseinheit gezogen wurde (vgl. hierzu auch Bemerkung 1.3.2).

Es ist bekannt, daß man θ_1^2 im allgemeinen nicht erwartungstreu durch $(\theta_1^*)^2$ schätzen kann. Falls nun $\text{var}(T_1)$ bekannt ist oder erwartungstreu durch $\widehat{\text{var}}(T_1)$ geschätzt werden kann, ist es auch möglich, θ_1^2 erwartungstreu zu schätzen durch (vgl. Gl. (1.3.16)):

$$(1.5.6) \quad (\theta_1^2)^* = a_{11}^* + a_{12}^* - \text{var}(T_1) \quad \text{bzw.}$$

$$(1.5.7) \quad (\theta_1^2)^* = a_{11}^* + a_{12}^* - \widehat{\text{var}}(T_1)$$

Um eine Schätzung des Risikos einer beliebigen linearen, symmetrischen Schätzfunktion zu erhalten, braucht man jetzt nur in Gl. (1.3.18) oder (1.3.4) die Schätzwerte bzw. die bekannten Parameter einzusetzen. Speziell gilt für die Schätzung der Varianz von θ_1^* (vgl. Gl. (1.3.24)):

$$(1.5.8) \quad \widehat{\text{var}}(\theta_1^*) = \frac{n^3}{n-1} \sigma_1^{2*} \left(E\left(\frac{1}{K}\right) - \frac{1}{n} \right) + \text{var}(T_1) \\ = \left(E\left(\frac{1}{K}\right) - \frac{1}{n} \right) \frac{n^2}{K-1} \left(\sum_{i=1}^n \left(x_i - \frac{1}{K} \sum_{j=1}^n x_j s_j \right)^2 s_i \right) + \text{var}(T_1),$$

wobei für $\text{var}(T_1)$ gegebenenfalls eine erwartungstreue Schätzfunktion eingesetzt wird.

Für die Schätzfunktion $\frac{1}{n}\theta_1^*$ des Mittels $\frac{1}{n}\theta_1$ beim k-maligen Ziehen ohne Zurücklegen für feste Werte $X_i \equiv x_i$ erhält man die bekannte Formel (vgl. Stenger (1971) S. 74):

$$(1.5.9) \quad \widehat{\text{var}}\left(\frac{1}{n}\theta_1^*\right) = \left(1 - \frac{k}{n}\right) \frac{1}{n} \frac{1}{k-1} \sum_{i=1}^n \left(x_i - \frac{1}{k} \sum_{j=1}^n x_j s_j \right)^2 s_i.$$

Es bleibt nun die Aufgabe, Aussagen über $\text{var}(T_1)$ zu gewinnen. Im Normalfall der Stichprobentheorie werden die Werte in der Grundgesamtheit nicht als Realisierungen von Zufallsvariablen aufgefaßt und wir wissen dann, daß $\text{var}(T_1)=0$ gilt. Wenn die Werte $X_i, i=1, \dots, n$ unabhängige, identisch verteilte Zufallsvariablen sind, gilt $\sigma_1^2 = \frac{n-1}{n} \text{var}(T_1)$ und aus der Schätzung von σ_1^2 kann man eine Schätzung von $\text{var}(T_1)$ herleiten. Im allgemeinen hat man allerdings bei dem bisherigen Modell keine Information über $\text{var}(T_1)$: von T_1 wird keine Realisierung beobachtet und von den Zufallsvariablen X_i wird in der

Stichprobe jeweils nur eine Realisierung beobachtet; wir können aus dieser Realisierung im allgemeinen weder $\text{var}(X_i)$ noch $\text{cov}(X_i, X_j)$ schätzen. Bei Vorliegen von verschiedenen Realisierungen x_i und x_j von zwei Untersuchungseinheiten haben wir ohne zusätzliche Information keine Möglichkeit zu unterscheiden, ob es sich hier z.B. um Realisierungen von identisch verteilten, unabhängigen Zufallsvariablen oder um Realisierungen von Zufallsvariablen X_i, X_j mit $E(X_i)=x_i$, $E(X_j)=x_j$ und $\text{var}(X_i)=\text{var}(X_j)=0$ handelt. Eine naheliegende Möglichkeit, Informationen aus der Stichprobe über $\text{var}(X_i)$ und $\text{cov}(X_i, X_j)$ zu bekommen, besteht darin, zu den Untersuchungseinheiten der Stichprobe S mehrere Realisierungen der Zufallsvariablen X_i zu beobachten, wobei wir nur den Fall untersuchen wollen, bei dem diese Beobachtungen stochastisch unabhängig voneinander sind. $X_{iv}, v=1, \dots, \kappa$ seien die zur v -ten Beobachtung der i -ten Untersuchungseinheit gehörigen Zufallsvariablen und die Mengen von Zufallsvariablen $\{X_{iv}, i \in \Omega\} v=1, \dots, \kappa$ seien stochastisch unabhängig voneinander. Außerdem seien für jedes $i \in \Omega$ die Zufallsvariablen $X_{iv}, v=1, \dots, \kappa$ identisch verteilt. Wir betrachten die Zufallsvariablen:

$$(1.5.10) \quad X_i = \frac{1}{\kappa} \sum_{v=1}^{\kappa} X_{iv} .$$

Für i aus der Stichprobe S werden alle $X_{iv}, v=1, \dots, \kappa$ beobachtet.

Die bisherigen Ergebnisse liefern Schätzfunktionen für

$\theta_1 = E\left(\sum_{i=1}^n X_i\right) = E\left(\sum_{i=1}^n \frac{1}{\kappa} \sum_{v=1}^{\kappa} X_{iv}\right)$ und für die Streuung σ_1^2 der X_i aus einer symmetrischen Stichprobe aus der Menge Ω der Untersuchungseinheiten. Die Varianz der X_i können wir für $i \in S$ durch

$$(1.5.11) \quad \sigma_i^{2*} = \frac{1}{\kappa(\kappa-1)} \sum_{v=1}^{\kappa} (X_{iv} - X_i)^2$$

erwartungstreu schätzen, die Kovarianz von zwei Zufallsvariablen X_i und $X_j, i, j \in S, i \neq j$ durch:

$$(1.5.12) \quad \sigma_{i,j}^* = \frac{1}{K(K-1)} \sum_{v=1}^K (X_{iv} - X_i)(X_{jv} - X_j)$$

(vgl. Stenger (1971) S. 31,33).

Als Schätzfunktionen V_1^* für $\sum_{i=1}^n \text{var}(X_i)$ und C_1^* für $\sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \text{cov}(X_i, X_j)$ erhalten wir dann:

$$(1.5.13) \quad V_1^* = \frac{n}{K} \sum_{i=1}^n \sigma_i^{2*} S_i \quad \text{und}$$

$$(1.5.14) \quad C_1^* = \frac{n(n-1)}{K(K-1)} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \sigma_{ij}^* S_i S_j \quad \text{mit}$$

$$(1.5.15) \quad E(V_1^*) = \sum_{i=1}^n E(\sigma_i^{2*}) E\left(\frac{n}{K} S_i\right) = \sum_{i=1}^n \text{var}(X_i)$$

$$(1.5.16) \quad E(C_1^*) = \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n E(\sigma_{ij}^*) E\left(\frac{n(n-1)}{K(K-1)} S_i S_j\right) = \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \text{cov}(X_i, X_j)$$

Wegen $\text{var}(T_1) = \sum_{i=1}^n \text{var}(X_i) + \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \text{cov}(X_i, X_j)$ erhalten wir als

erwartungstreue Schätzfunktion $\text{var}(T_1)^*$ für $\text{var}(T_1)$:

$$(1.5.17) \quad \text{var}(T_1)^* = V_1^* + C_1^* .$$

Die Schätzfunktionen V_1^* und C_1^* sind als Schätzung der Summe der Varianzen und der Abhängigkeit der Werte auch für sich interessant. Außerdem ermöglichen sie es, die Streuung σ_{11}^2 der Erwartungswerte (vgl. Gl. (1.3.16)) der X_i wie folgt zu schätzen:

$$(1.5.18) \quad \sigma_{11}^{2*} = \sigma_1^{2*} - \frac{n-1}{n^2} V_1^* + \frac{1}{n^2} C_1^*$$

Die hier durchgeführte Herleitung weist auf eine wichtige Anwendungsmöglichkeit der Auffassung der Werte als Zufallsvariablen hin: Zu jeder Untersuchungseinheit bzw. jedem Paar von Untersuchungseinheiten der Stichprobe wurden nicht feste Werte, sondern Realisierungen der Schätzfunktionen X_i für $E(X_{iv})$, σ_i^{2*} und σ_{ij}^* beobachtet (vgl. Abschnitt 3.2.5).

II. SCHÄTZPROBLEME BEI ZUFALLSVARIABLEN, DIE
PAAREN VON UNTERSUCHUNGSEINHEITEN ZUGEORDNET
SIND

2.1 Problemstellung, Voraussetzungen, Bezeichnungen

In Kapitel I waren den Untersuchungseinheiten i der Grundgesamtheit Ω Zufallsvariablen X_i zugeordnet; wir wollen nun den Fall betrachten, daß Paaren $\{i, j\}$, $i \neq j$ von Untersuchungseinheiten Zufallsvariablen $\bar{X}_{ij}$ zugeordnet sind. Da $\{i, j\}$ und $\{j, i\}$ das gleiche Paar bezeichnen soll, gilt $\bar{X}_{ij} = \bar{X}_{ji}$ und man kann bei der Indizierung $j > i$ wählen. Anstelle der ursprünglichen n -elementigen Grundgesamtheit Ω können wir nun die Menge aller Paare $\Omega_2 := \{\{i, j\} : i, j \in \Omega, j > i\}$ betrachten und der zugehörige Vektor von Zufallsvariablen $(\bar{X}_{12}, \dots, \bar{X}_{ij}, \dots, \bar{X}_{(n-1)n})$ habe eine Verteilung W aus einer Menge $\mathcal{W}$ von $\frac{1}{2}n(n-1)$ dimensionalen Verteilungen. Aus einer Stichprobe soll der Erwartungswert θ_2 der Summe T_2 dieser Zufallsvariablen geschätzt werden.

$$(2.1.1) \quad \theta_2 = E(T_2) = E\left(\sum_{i=1}^n \sum_{j>i}^n \bar{X}_{ij}\right)$$

Man hat also soweit das gleiche Problem wie in Abschnitt 1.2 mit dem nur formalen Unterschied, daß die Untersuchungseinheiten und Zufallsvariablen doppelt indiziert sind. Das Stichprobenverfahren wird aus symmetrischen Stichprobenverfahren bzgl. Ω konstruiert: Zu einer Stichprobe $Sc\Omega$ werden die Werte der Paare $\{i, j\}$ dann beobachtet, wenn i und j aus S sind.

Voraussetzung 2.1.1

Für die Stichprobenindikatorvariablen S_{ij} der Paare und S_i, S_j der Untersuchungseinheiten gelte:

$$(2.1.2) \quad S_{ij} = S_i \cdot S_j = \begin{cases} 1 & \text{falls } i \in S \text{ und } j \in S \\ 0 & \text{sonst} \end{cases}$$

Das Stichprobenverfahren P bezüglich der Grundgesamtheit Ω sei symmetrisch mit:

$$(2.1.3) \quad \alpha_0 + \alpha_1 = P(K \leq 1) = 0$$

Außerdem gelte $n = |\Omega| \geq 4$.

Es sei angemerkt, daß das konstruierte Stichprobenverfahren bezüglich der Gesamtheit Ω_2 nicht symmetrisch ist, da z.B. $E(S_{ij}S_{il}) \neq E(S_{ij}S_{ml})$ für $m \notin \{i, j, l\}$ gilt. Gleichung (2.1.3) soll sicherstellen, daß in der Stichprobe mit Wahrscheinlichkeit 1 der Wert eines Paares beobachtet wird. Die Annahme $n \geq 4$ wurde gemacht, um später notwendige Sonderbetrachtungen für die praktisch unwichtigen Fälle $n \leq 3$ zu vermeiden. Falls die Stichprobe $S \subset \Omega$ k Elemente hat, wird der Wert von $\frac{1}{2}k \cdot (k-1)$ Zufallsvariablen beobachtet; es gilt also

$$(2.1.4) \quad \sum_{i=1}^n \sum_{j>i}^n S_{ij} = \frac{1}{2}K(K-1),$$

und der Umfang der Stichprobe aus Ω_2 ist eine Funktion von K .

Voraussetzung 2.1.2

Behandelt werden nur lineare, symmetrische Schätzfunktionen, d.h. Schätzfunktion $\hat{\theta}_2$ der Form:

$$(2.1.5) \quad \hat{\theta}_2 = \gamma(K) \sum_{i=1}^n \sum_{j>i}^n \bar{X}_{ij} S_i S_j .$$

Im folgenden werden wir auch die Kovarianz von $\hat{\theta}_1$ und $\hat{\theta}_2$ angeben, also Werte von Untersuchungseinheiten und von Paaren von Untersuchungseinheiten simultan betrachten. Wir können dann von einer Grundgesamtheit $\Omega_3 = \{\{i,j\} : i \in \Omega, j \in \Omega, j \geq i\}$ ausgehen und die zugehörigen Zufallsvariablen $(\bar{X}_{11}, \dots, \bar{X}_{ij}, \dots, \bar{X}_{nn})$ mit $\bar{X}_{ii} = X_i$ haben eine unbekannte Verteilung W aus einer Menge von $\frac{1}{2}n(n+1)$ dimensionalen Verteilungen. Die Voraussetzung über die Unabhängigkeit der Werte vom Stichprobenverfahren soll für diesen umfassenderen Fall formuliert werden.

Voraussetzung 2.1.3

Die Familie von Zufallsvariablen $\{\bar{X}_{ij} : i \in \Omega, j \in \Omega\}$ sei von der Menge von Zufallsvariablen $\{S_i : i \in \Omega\}$ unabhängig.

Funktionen $f(\bar{X}_{11}, \dots, \bar{X}_{ij}, \dots, \bar{X}_{nn})$ und $g(S_1, \dots, S_n)$ sind dann auch unabhängig voneinander (vgl. Krickeberg (1962) S. 76) und insbesondere gilt für die Erwartungswerte:

$$(2.1.6) \quad \begin{aligned} E(f(\bar{X}_{11}, \dots, \bar{X}_{nn})g(S_1, \dots, S_n)) \\ = E(f(\bar{X}_{11}, \dots, \bar{X}_{nn}))E(g(S_1, \dots, S_n)). \end{aligned}$$

Nebenbei sei angemerkt, daß man die Schätzung von θ_2 als Schätzung des Erwartungswertes einer Summe von Zufallsvariablen ansehen kann, die sich auf die Untersuchungseinheiten beziehen, wenn $X_i = \frac{1}{2} \sum_{\substack{j=1 \\ j \neq i}}^n \bar{X}_{ij}$ gesetzt wird. Zu

Untersuchungseinheiten $i \in S$ werden dann allerdings nicht

Realisierungen von X_i , sondern von $\sum_{\substack{j=1 \\ j \neq i}}^n \bar{X}_{ij} S_j$ beobachtet

und somit sind im allgemeinen die beobachteten Werte von den Stichprobenindikatorvariablen abhängig.

Bemerkung 2.1.4

Mit dem beschriebenen Modell kann auch der Fall behandelt werden, daß gerichteten Tupeln $(i,j), i,j \in \Omega, j \neq i$ Variablen X_{ij} zugeordnet werden und so $X_{ij} \neq X_{ji}$ gelten kann. Für $\bar{X}_{ij} := X_{ij} + X_{ji}$ gilt:

$$(2.1.7) \quad \theta_2 = \sum_{i=1}^n \sum_{j>i}^n \bar{X}_{ij} = \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n X_{ij} \quad \text{und}$$

$$(2.1.8) \quad \hat{\theta}_2 = \gamma(K) \sum_{i=1}^n \sum_{j>i}^n \bar{X}_{ij} S_i S_j = \gamma(K) \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n X_{ij} S_i S_j ;$$

man braucht also nur die Zufallsvariablen $\bar{X}_{ij}$ der ungerichteten Paare $\{i,j\}$ zu betrachten.

In diesem Kapitel werden wir Eigenschaften von linearen, symmetrischen Schätzfunktionen $\hat{\theta}_2$ untersuchen und analog zu Kapitel I Schätzfunktionen für andere Parameter der Verteilung von $(\bar{X}_{12}, \dots, \bar{X}_{ij}, \dots, \bar{X}_{(n-1)n})$ angeben. Zur simultanen Schätzung von θ_1 und θ_2 oder zur Schätzung von möglicherweise gewichteten Summen von θ_1 und θ_2 wird $\text{cov}(\hat{\theta}_1, \hat{\theta}_2)$ berechnet und geschätzt.

Wenn wir $\bar{X}_{ij} = X_i \cdot X_j$ setzen, erhalten wir gegenüber Abschnitt 1.5 weitere Aussagen zur Schätzung der Parameter

a_{12}, σ_1^2 usw.. Außerdem wird die Grundlage für die in Kapitel III behandelte Schätzung von Kantensummen in Graphen geschaffen.

Am Schluß dieses Abschnittes soll noch die Bezeichnungsweise erklärt werden: Zufallsvariablen und Parameter, die sich auf die einzelnen Untersuchungseinheiten, also auf X_i , $i \in \Omega$ beziehen, erhalten als ersten Index 1, z.B. $\theta_1, \hat{\theta}_1, T_1, \sigma_1^2, a_{11}, a_{12}$ in Kapitel I. Wenn sie sich auf Paare und die Zufallsvariablen $\bar{X}_{ij}$, $i, j \in \Omega, j > i$ beziehen, erhalten sie den ersten Index 2, z.B. $\theta_2, \hat{\theta}_2, T_2$. Wenn sie sich auf $(\bar{X}_{11}, \bar{X}_{12}, \dots, \bar{X}_{ij}, \dots, \bar{X}_{nn})$ beziehen, werden sie mit dem Index 3 versehen, z.B. $\theta_3 = \theta_1 + \theta_2$.

Der Balken bei $\bar{X}_{ij}$ soll zeigen, daß die Variable einem Paar $\{i, j\}$ und nicht einem gerichteten Tupel (i, j) zugeordnet ist und somit $\bar{X}_{ij} = \bar{X}_{ji}$ gilt.

Wie im ersten Kapitel beziehen sich Erwartungswerte, Wahrscheinlichkeiten und Parameter auf eine Verteilung, die von einer unbekannten, festen Verteilung W der Werte und einem festgelegten Stichprobenverfahren P erzeugt wird. Wahrscheinlichkeiten bezüglich dieser Verteilung werden mit $P(\cdot)$ z.B. $P(\hat{\theta}_2 > 0)$ angegeben; an dem Argument ist zu erkennen, daß der Buchstabe P in diesen Fällen nicht immer das Stichprobenverfahren bezeichnet.

In diesem Kapitel gelten stets alle hier genannten Voraussetzungen.

2.2 Die Schätzung des Erwartungswertes der Summe der Zufallsvariablen

Zunächst geben wir $E(\hat{\theta}_2 - \theta_2)^2$ mit Parametern an, die denen von Satz 1.3.3 entsprechen.

Satz 2.2.1

Es gilt:

$$(2.2.1) \quad E(\hat{\theta}_2) = \theta_2 E\left(\gamma(K) \frac{K(K-1)}{n(n-1)}\right) \text{ und}$$

$$(2.2.2) \quad E(\hat{\theta}_2 - \theta_2)^2 = a_{22} E\left(\gamma^2(K) \frac{K(K-1)}{n(n-1)}\right) \\ + a_{23} E\left(\gamma^2(K) \frac{K(K-1)(K-2)}{n(n-1)(n-2)}\right) \\ + a_{24} E\left(\gamma^2(K) \frac{K(K-1)(K-2)(K-3)}{n(n-1)(n-2)(n-3)}\right) \\ - \theta_2^2 \left(2E\left(\gamma(K) \frac{K(K-1)}{n(n-1)}\right) - 1\right)$$

mit

$$(2.2.3) \quad a_{22} := \sum_{i=1}^n \sum_{j>i}^n E(\bar{X}_{ij}^2)$$

$$(2.2.4) \quad a_{23} := \sum_{i=1}^n \sum_{j>i}^n \sum_{\substack{\ell=1 \\ \ell \neq i \\ \ell \neq j}}^n E(\bar{X}_{ij} (\bar{X}_{i\ell} + \bar{X}_{j\ell}))$$

$$(2.2.5) \quad a_{24} := \sum_{i=1}^n \sum_{j>i}^n \sum_{\substack{\ell=1 \\ \ell \neq i \\ \ell \neq j}}^n \sum_{\substack{m>\ell \\ m \neq i \\ m \neq j}}^n E(\bar{X}_{ij} \bar{X}_{\ell m})$$

Bemerkung 2.2.2

Vor dem Beweis des Satzes sollen die Parameter erklärt werden. Der erste Index bei a_{22} , a_{23} , a_{24} weist darauf hin, daß die Parameter sich auf Paare von Untersuchungseinheiten beziehen. Die Parameter sind Summen der Erwartungswerte von Produkten $X_{ij}X_{lm}$ und die Anzahl $|\{i,j,\ell,m\}|$ der Untersuchungseinheiten, auf die sich diese Produkte beziehen, ist jeweils der zweite Index dieser Parameter.

In a_{23} und a_{24} tauchen wegen $X_{ij}X_{lm} = X_{lm}X_{ij}$ Summanden doppelt auf und wir können die Parameter anders schreiben:

$$(2.2.6) \quad a_{23} = E\left(\sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \sum_{\substack{\ell=1 \\ \ell \neq i \\ \ell \neq j}}^n \bar{X}_{ij} \bar{X}_{i\ell}\right) = 2E\left(\sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \sum_{\substack{\ell > j \\ \ell \neq i}}^n \bar{X}_{ij} \bar{X}_{i\ell}\right)$$

$$(2.2.7) \quad a_{24} = 2E\left(\sum_{i=1}^n \sum_{j>i}^n \sum_{\substack{\ell>i \\ \ell \neq j}}^n \sum_{\substack{m>\ell \\ m \neq j}}^n \bar{X}_{ij} \bar{X}_{lm}\right),$$

Summiert wird nun bei a_{23} über alle (ungeordneten) verschiedenen Paare $\{(i,j),(\ell,m)\}$, die genau eine Untersuchungseinheit gemeinsam haben (d.h. $|\{i,j,\ell,m\}|=3$) und bei a_{24} über alle Paare, die keine Untersuchungseinheit gemeinsam haben. Die Parameter sind das Doppelte des Erwartungswertes der Summe der zugehörigen Zufallsvariablen $X_{ij}X_{lm}$. Frank (1977a) gibt noch andere Darstellungen der hier genannten Parameter an.

Bei a_{23} können wir anstelle der Produkte $X_{ij}X_{i\ell}$ die Zufallsvariablen betrachten:

$$(2.2.8) \quad U_i = \sum_{\substack{j=1 \\ j \neq i}}^n \bar{X}_{ij}$$

Es gilt:

$$(2.2.9) \quad \begin{aligned} a_{23} &= \sum_{i=1}^n E \left[\left(\sum_{\substack{j=1 \\ j \neq i}}^n \bar{X}_{ij} \right) \left(\sum_{\substack{\ell=1 \\ \ell \neq i}}^n \bar{X}_{i\ell} \right) \right] - \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n E(\bar{X}_{ij}^2) \\ &= \sum_{i=1}^n E(U_i^2) - 2a_{22} \end{aligned}$$

Die folgende Beziehung werden wir oft benützen:

$$(2.2.10) \quad \begin{aligned} E(T_2^2) &= \text{var}(T_2) + \theta_2^2 \\ &= a_{22} + a_{23} + a_{24} \end{aligned}$$

Nach dieser Gleichung können wir den Parameter a_{24} durch $E(T_2^2)$ oder $\text{var}(T_2)$ ersetzen.

Beweis von Satz 2.2.1

Nach Gleichung (1.1.28) und wegen der Unabhängigkeit der Werte vom Stichprobenverfahren gilt:

$$(2.2.11) \quad \begin{aligned} E(\hat{\theta}_2) &= \sum_{i=1}^n \sum_{j>i}^n \bar{X}_{ij} E(\gamma(K) S_i S_j) \\ &= \theta_2 E\left(\gamma(K) \frac{K(K-1)}{n(n-1)}\right) \end{aligned}$$

$$(2.2.12) \quad \begin{aligned} E(\hat{\theta}_2 - \theta_2)^2 &= E(\hat{\theta}_2^2) - 2\theta_2 E(\hat{\theta}_2) + \theta_2^2 \\ &= \sum_{i=1}^n \sum_{j>i}^n \sum_{\ell=1}^n \sum_{m>\ell}^n E(X_{ij} X_{\ell m}) E(\gamma^2(K) S_i S_j S_\ell S_m) \\ &\quad - \theta_2^2 \left(2E\left(\gamma(K) \frac{K(K-1)}{n(n-1)}\right) - 1 \right) \end{aligned}$$

$$\begin{aligned}
 &= \sum_{i=1}^n \sum_{j>i}^n E(\bar{X}_{ij}^2) E(\gamma^2(K) \frac{K(K-1)}{n(n-1)}) \\
 &+ \left(\sum_{i=1}^n \sum_{j>i}^n \sum_{\substack{\ell<i \\ \ell \neq j}}^n E(\bar{X}_{ij} \bar{X}_{\ell i}) + \sum_{i=1}^n \sum_{j>i}^n \sum_{\substack{m>i \\ m \neq j}}^n E(\bar{X}_{ij} \bar{X}_{im}) \right. \\
 &+ \sum_{i=1}^n \sum_{j>i}^n \sum_{\substack{\ell<j \\ \ell \neq i}}^n E(\bar{X}_{ij} \bar{X}_{\ell j}) + \left. \sum_{i=1}^n \sum_{j>i}^n \sum_{m>j}^n E(\bar{X}_{ij} \bar{X}_{jm}) \right) \\
 &E(\gamma^2(K) \frac{K(K-1)(K-2)}{n(n-1)(n-2)}) \\
 &+ \sum_{i=1}^n \sum_{j>i}^n \sum_{\substack{\ell=1 \\ \ell \neq i \\ \ell \neq j}}^n \sum_{\substack{m>\ell \\ m \neq i \\ m \neq j}}^n E(\bar{X}_{ij} \bar{X}_{\ell m}) E(\gamma^2(K) \frac{K(K-1)(K-2)(K-3)}{n(n-1)(n-2)(n-3)}) \\
 &- \theta_2^2 \left(2E(\gamma(K) \frac{K(K-1)}{n(n-1)}) - 1 \right) .
 \end{aligned}$$

Da wegen $\bar{X}_{ij} = \bar{X}_{ji}$ gilt:

$$\begin{aligned}
 (2.2.13) \quad & \sum_{i=1}^n \sum_{j>i}^n \sum_{\ell<i}^n E(\bar{X}_{ij} \bar{X}_{\ell i}) + \sum_{i=1}^n \sum_{j>i}^n \sum_{\substack{m>i \\ m \neq j}}^n E(\bar{X}_{ij} \bar{X}_{im}) \\
 &+ \sum_{i=1}^n \sum_{j>i}^n \sum_{\substack{\ell<j \\ \ell \neq i}}^n E(\bar{X}_{ij} \bar{X}_{\ell j}) + \sum_{i=1}^n \sum_{j>i}^n \sum_{m>j}^n E(\bar{X}_{ij} \bar{X}_{jm}) \\
 &= \sum_{i=1}^n \sum_{j>i}^n \left(\sum_{\substack{\ell=1 \\ \ell \neq j \\ \ell \neq i}}^n E(\bar{X}_{ij} \bar{X}_{i\ell}) + \sum_{\substack{\ell=1 \\ \ell \neq j \\ \ell \neq i}}^n E(\bar{X}_{ij} \bar{X}_{j\ell}) \right) = a_{23},
 \end{aligned}$$

folgt die Gültigkeit von Gl. (2.2.2).

Bemerkung 2.2.3

Lineare und symmetrische Schätzfunktionen von θ_2 sind nach Satz 2.2.1 genau dann erwartungstreu, wenn gilt:

$$(2.2.14) \quad E\left(\gamma(K) \frac{K(K-1)}{n(n-1)}\right) = 1.$$

Wir werden wieder insbesondere die erwartungstreuen Schätzfunktionen mit $\gamma(K) = \frac{n(n-1)}{K(K-1)}$ und $\gamma(K) = \frac{n(n-1)}{E(K(K-1))}$ betrachten und erhalten im ersten Fall das $\frac{n(n-1)}{2}$ -fache des arithmetischen Mittels der beobachteten $\frac{K(K-1)}{2}$ Werte:

$$(2.2.15) \quad \theta_2^* := \frac{n(n-1)}{K(K-1)} \sum_{i=1}^n \sum_{j>i}^n \bar{X}_{ij} S_{ij}$$

θ_2^* ist nur definiert, falls $\alpha_0 + \alpha_1 = P(K \leq 1) = 0$.

Im zweiten Fall erhalten wir die Horvitz-Thompson-Schätzfunktionen:

$$(2.2.16) \quad \begin{aligned} \tilde{\theta}_2 &:= \frac{n(n-1)}{E(K(K-1))} \sum_{i=1}^n \sum_{j>i}^n \bar{X}_{ij} S_{ij} \\ &= \frac{1}{p_2} \sum_{i=1}^n \sum_{j>i}^n \bar{X}_{ij} S_{ij} \end{aligned}$$

Für die quadratische Abweichung der Schätzfunktionen vom wahren Wert θ_2 gilt:

$$(2.2.17) \quad \begin{aligned} E(\theta_2^* - \theta_2)^2 &= \text{var}(\theta_2^*) \\ &= a_{22} n(n-1) E\left(\frac{1}{K(K-1)}\right) \end{aligned}$$

$$+ a_{23} \frac{n(n-1)}{n-2} E\left(\frac{1}{K} - \frac{1}{K(K-1)}\right) \\ + a_{24} \frac{n(n-1)}{(n-2)(n-3)} E\left(1 - \frac{4}{K} + \frac{2}{K(K-1)}\right) - \theta_2^2$$

$$(2.2.18) \quad E(\tilde{\theta}_2 - \theta_2)^2 = \text{var}(\tilde{\theta}_2)$$

$$= a_{22} \frac{1}{p_2} + a_{23} \frac{p_3}{p_2^2} + a_{24} \frac{p_4}{p_2^2} - \theta_2^2.$$

Wie in Abschnitt 1.3 sollen nun die Parameter a_{22} , a_{23} und a_{24} durch Maße für die Streuung der Werte ersetzt werden. Außer der Streuung der $\bar{X}_{ij}$ brauchen wir bei der Angabe von $E(\hat{\theta}_2 - \theta_2)^2$ einen weiteren Parameter. Im weiteren Verlauf erweist sich die Streuung der in Gl. (2.2.8) eingeführten Variablen U_i als sinnvoll: im Rahmen der Graphentheorie ist sie als Grad eines Knotens gut interpretierbar (vgl. Kap. III). Bei den Betrachtungen zur Zulässigkeit (Abschnitt 2.3) haben Streuungen als Parameter die wichtige Eigenschaft, nicht negativ zu sein. Für die Variablen U_i gilt:

$$(2.2.19) \quad E\left(\frac{1}{n} \sum_{i=1}^n U_i\right) = E\left(\frac{2}{n} T_2\right) = \frac{2}{n} \theta_2.$$

Analog zu Gl. (1.3.15) definieren wir als Streuung σ_2^2 der $\bar{X}_{ij}$ und als Streuung σ_u^2 der U_i

$$(2.2.20) \quad \sigma_2^2 := E\left(\frac{2}{n(n-1)} \sum_{i=1}^n \sum_{j>i}^n (\bar{X}_{ij} - \frac{2}{n(n-1)} \sum_{i=1}^n \sum_{j>i}^n \bar{X}_{ij})^2\right) \\ = \frac{2}{n(n-1)} a_{22} - \frac{4}{n^2(n-1)^2} E(T_2^2) \\ = \frac{2(n+1)(n-2)}{n^2(n-1)^2} a_{22} - \frac{4}{n^2(n-1)^2} (a_{23} + a_{24})$$

$$\begin{aligned}
 (2.2.21) \quad \sigma_u^2 &:= E\left(\frac{1}{n} \sum_{i=1}^n (U_i - \frac{1}{n} \sum_{i=1}^n U_i)^2\right) \\
 &= \frac{1}{n} \sum_{i=1}^n E(U_i^2) - \frac{4}{n^2} E(T_2^2) \\
 &= \frac{2}{n} a_{22} + \frac{1}{n} a_{23} - \frac{4}{n^2} E(T_2^2) \\
 &= \frac{2(n-2)}{n^2} a_{22} + \frac{n-4}{n^2} a_{23} - \frac{4}{n^2} a_{24}
 \end{aligned}$$

Für konstante Werte $\bar{X}_{ij} \equiv \bar{x}_{ij}$ handelt es sich um die in der Stichprobentheorie üblichen Kennzahlen für die Streuung der $u_i = \sum_{\substack{j=1 \\ j \neq i}}^n \bar{x}_{ij}$ und der $\bar{x}_{ij}$.

Bemerkung 2.2.4

Mit Hilfe der Schwarzschen Ungleichung (siehe Feller (1971) S. 152) läßt sich die Beziehung zwischen σ_u^2 und σ_2^2 klären:

$$\begin{aligned}
 (2.2.22) \quad \sigma_u^4 &= \frac{1}{n^2} \left(\sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n E\left((U_i - \frac{2}{n} T_2)(\bar{X}_{ij} - \frac{2}{n(n-1)} T_2)\right) \right)^2 \\
 &\leq \frac{1}{n^2} \left(\sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n E(U_i - \frac{2}{n} T_2)^2 \right) \\
 &\quad \left(\sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n E(X_{ij} - \frac{2}{n(n-1)} T_2)^2 \right) \\
 &= (n-1)^2 \sigma_u^2 \sigma_2^2
 \end{aligned}$$

Daraus folgt:

$$(2.2.23) \quad \sigma_u^2 \leq (n-1)^2 \sigma_2^2$$

Die Berücksichtigung der Tatsache, daß σ_u^2 und σ_2^2 nicht unabhängig voneinander beliebige Werte annehmen können, ist in Abschnitt 2.3 bei den Betrachtungen zur Zulässigkeit wichtig. Außerdem spielt die obige Ungleichung in Abschnitt 3.7 eine große Rolle.

Satz 2.2.5

Es gilt:

$$(2.2.24) \quad E(\hat{\theta}_2 - \theta_2)^2 = \sigma_2^2 f_1(\gamma) + \sigma_u^2 f_2(\gamma) \\ + \text{var}(T_2) f_3(\gamma) + \theta_2^2 f_4(\gamma)$$

mit:

$$(2.2.25) \quad f_1(\gamma) := E\left(\gamma^2 (K) \frac{K(K-1)}{2} \left(1 - \frac{2(K-2)}{n-2} + \frac{(K-2)(K-3)}{(n-2)(n-3)}\right)\right) \\ = E\left(\gamma^2 (K) \frac{K(K-1)(n-K)(n-K-1)}{2(n-2)(n-3)}\right)$$

$$(2.2.26) \quad f_2(\gamma) := E\left(\gamma^2 (K) \frac{K(K-1)(K-2)}{(n-1)(n-2)} \left(1 - \frac{K-3}{n-3}\right)\right) \\ = E\left(\gamma^2 (K) \frac{K(K-1)(K-2)(n-K)}{(n-1)(n-2)(n-3)}\right)$$

$$(2.2.27) \quad f_3(\gamma) = E \left[\left(\gamma(K) \frac{K(K-1)}{n(n-1)} \right)^2 \right]$$

$$(2.2.28) \quad f_4(\gamma) = E \left[\left(\gamma(K) \frac{K(K-1)}{n(n-1)} - 1 \right)^2 \right]$$

Beweis

Aus den Gleichungen (2.2.10) (2.2.20) (2.2.21) erhält man durch Auflösen nach a_{22} , a_{23} und a_{24} :

$$(2.2.29) \quad a_{22} = \frac{n(n-1)}{2} \sigma_2^2 + \frac{2}{n(n-1)} \text{var}(T_2) + \frac{2}{n(n-1)} \theta_2^2$$

$$(2.2.30) \quad a_{23} = n \sigma_u^2 - n(n-1) \sigma_2^2 + \frac{4(n-2)}{n(n-1)} \text{var}(T_2) + \frac{4(n-2)}{n(n-1)} \theta_2^2$$

$$(2.2.31) \quad a_{24} = \frac{(n-2)(n-3)}{n(n-1)} \text{var}(T_2) + \frac{(n-2)(n-3)}{n(n-1)} \theta_2^2 \\ + \frac{n(n-1)}{2} \sigma_2^2 - n \sigma_u^2$$

Das Einsetzen dieser Gleichungen in Gl. (2.2.2) und das Zusammenfassen der Koeffizienten von σ_2^2 , σ_u^2 , $\text{var}(T_2)$ und θ_2^2 ergibt die Werte $f_1(\gamma)$, $f_2(\gamma)$, $f_3(\gamma)$ und $f_4(\gamma)$.

Bemerkung 2.2.6

Bei erwartungstreuen Schätzfunktionen (vgl. Gl. (2.2.14)) gilt für den Koeffizienten $f_4(\gamma)$ von θ_2^2 :

$$f_4(\gamma) = \text{var} \left(\gamma(K) \frac{K(K-1)}{n(n-1)} \right)$$

Speziell für θ_2^* (d.h. $\gamma(K) = \frac{n(n-1)}{K(K-1)}$) beträgt dieser Faktor 0. Wie im ersten Kapitel (vgl. Bem. 1.3.7) hängt $E(\theta_2^* - \theta_2)^2$ nicht von θ_2 ab. Die Koeffizienten der anderen Parameter lauten für θ_2^* :

$$(2.2.32) \quad f_1(\gamma) = \frac{n^2(n-1)^2}{2(n-2)(n-3)} E\left(\frac{(n-K)(n-K-1)}{K(K-1)}\right)$$

$$(2.2.33) \quad f_2(\gamma) = \frac{n^2(n-1)}{(n-2)(n-3)} E\left(\frac{(K-2)(n-K)}{K(K-1)}\right)$$

$$(2.2.34) \quad f_3(\gamma) = 1$$

Bemerkung 2.2.7

Für die ebenfalls erwartungstreuen Horvitz-Thompson-Schätzfunktionen können die $f_i(\gamma)$ $i=1,2,3,4$ mit den Wahrscheinlichkeiten p_1, p_2, p_3, p_4 (Def. siehe Gl. (1.1.3)) angegeben werden. Wegen $\gamma(K) = \frac{n(n-1)}{E(K(K-1))} = \frac{1}{p_2}$ und nach Gl. (1.1.31) gilt:

$$(2.2.35) \quad f_1(\gamma) = \frac{n(n-1)}{2} \frac{p_2^{-2p_3+p_4}}{p_2^2}$$

$$(2.2.36) \quad f_2(\gamma) = n \frac{p_3^{-p_4}}{p_2^2}$$

$$(2.2.37) \quad f_3(\gamma) = \frac{(n-2)(n-3)}{n(n-1)} \frac{p_4}{p_2^2} + \frac{4(n-2)}{n(n-1)} \frac{p_3}{p_2^2} + \frac{2}{n(n-1)p_2}$$

$$(2.2.38) \quad f_4(\gamma) = f_3(\gamma) - 1$$

Die vorletzte Gleichung kann man mittels Gl. (1.1.31) und der Beziehung

$$(2.2.39) \quad K^2(K-1)^2 = K(K-1)(K-2)(K-3) + 4K(K-1)(K-2) + 2K(K-1)$$

berechnen.

Bemerkung 2.2.8

In Bemerkung 1.3.9 wurde das Modell der Superpopulation mit der üblichen Betrachtungsweise der Werte als Konstanten verglichen. Für gleiche Parameter σ_2^2 , σ_u^2 und θ_2^2 ist $E(\hat{\theta}_2 - \theta_2)^2$ für Zufallsvariablen $\bar{x}_{ij}$ um $\text{var}(T_2)f_3(\gamma)$ größer als bei ihrer Auffassung als feste Werte. Wiederum ist die Bedingung $\text{var}(T_2)=0$ ausreichend dafür, daß die unterschiedlichen Modellvorstellungen keine Auswirkungen haben. Auf den Erwartungswert von $\hat{\theta}_2$ haben beide Ansätze keinen Einfluß.

Bemerkung 2.2.9

Die Parameter σ_2^2 und σ_u^2 unterscheiden nicht, ob die Streuung durch unterschiedliche Erwartungswerte der $\bar{x}_{ij}$ bzw. U_i

oder durch die Varianzen der Zufallsvariablen erzeugt wird. Aus diesem Grunde sollen die Streuungen der entsprechenden Erwartungswerte definiert und zu σ_2^2 bzw. σ_u^2 in Beziehung gesetzt werden. Zunächst ist dazu eine Aufteilung von $\text{var}(T_2)$ notwendig:

$$\begin{aligned}
 (2.2.40) \quad \text{var}(T_2) &= E(T_2^2) - \theta_2^2 = a_{22} + a_{23} + a_{24} - \theta_2^2 \\
 &= a_{22} - \sum_{i=1}^n \sum_{j>i}^n E^2(\bar{X}_{ij}) \\
 &\quad + a_{23} - \sum_{i=1}^n \sum_{j>i}^n \sum_{\substack{\ell=1 \\ \ell \neq i \\ \ell \neq j}}^n E(\bar{X}_{ij}) E(\bar{X}_{ij} + \bar{X}_{j\ell}) \\
 &\quad + a_{24} - \sum_{i=1}^n \sum_{j>i}^n \sum_{\substack{\ell=1 \\ \ell \neq i \\ \ell \neq j}}^n \sum_{\substack{m>\ell \\ m \neq i \\ m \neq j}}^n E(\bar{X}_{ij}) E(\bar{X}_{\ell m}) \\
 &= c_{22} + c_{23} + c_{24}
 \end{aligned}$$

mit:

$$(2.2.41) \quad c_{22} := \sum_{i=1}^n \sum_{j>i}^n \text{var}(\bar{X}_{ij})$$

$$(2.2.42) \quad c_{23} := \sum_{i=1}^n \sum_{j>i}^n \sum_{\substack{\ell=1 \\ \ell \neq i \\ \ell \neq j}}^n \text{cov}(\bar{X}_{ij}, \bar{X}_{ij} + \bar{X}_{j\ell})$$

$$(2.2.43) \quad c_{24} := \sum_{i=1}^n \sum_{j>i}^n \sum_{\substack{\ell=1 \\ \ell \neq i \\ \ell \neq j}}^n \sum_{\substack{m>\ell \\ m \neq i \\ m \neq j}}^n \text{cov}(\bar{X}_{ij}, \bar{X}_{\ell m})$$

(zum Sinn der Indizierung vgl. Bem. 2.2.2)

Die Streuungen σ_{21}^2 von $E(\bar{X}_{ij})$ und σ_{u1}^2 von $E(U_i)$ lauten:

$$\begin{aligned} (2.2.44) \quad \sigma_{21}^2 &:= \frac{2}{n(n-1)} \sum_{i=1}^n \sum_{j>i}^n (E(\bar{X}_{ij}) - \frac{2}{n(n-1)}\theta)^2 \\ &= \frac{2}{n(n-1)} a_{22} - \frac{2}{n(n-1)} c_{22} - \frac{4}{n^2(n-1)^2} \theta_2^2 \\ &= \sigma_2^2 + \frac{4}{n^2(n-1)^2} \text{var}(T_2) - \frac{2}{n(n-1)} c_{22} \\ &= \sigma_2^2 - \frac{2(n+1)(n-2)}{n^2(n-1)^2} c_{22} \\ &\quad + \frac{4}{n^2(n-1)^2} (c_{23} + c_{24}) \end{aligned}$$

$$\begin{aligned} (2.2.45) \quad \sigma_{u1}^2 &= \frac{1}{n} \sum_{i=1}^n (E(U_i) - \frac{2}{n}\theta_2)^2 \\ &= \frac{1}{n} \sum_{i=1}^n E(U_i^2) - \frac{1}{n} \sum_{i=1}^n \text{var}(U_i) - \frac{4}{n^2} \theta_2^2 \\ &= \sigma_u^2 + \frac{4}{n^2} \text{var}(T_2) - \frac{1}{n} \sum_{i=1}^n \text{var}(U_i) \\ &= \sigma_u^2 - \frac{2(n-2)}{n^2} c_{22} - \frac{n-4}{n^2} c_{23} + \frac{4}{n^2} c_{24} \end{aligned}$$

Das letzte Gleichheitszeichen folgt hierbei aus:

$$(2.2.46) \quad \sum_{i=1}^n \text{var}(U_i) = 2 c_{22} + c_{23}$$

Mittels der hier genannten Gleichungen kann man bei der Angabe von $E(\hat{\theta}_2 - \theta_2)^2$ die Parameter σ_2^2 , σ_u^2 und $\text{var}(T_2)$ durch σ_{21}^2 , σ_{u1}^2 , c_{22} , c_{23} , c_{24} ersetzen, was wir allerdings nicht durchführen wollen.

Bemerkung 2.2.10

Abschließend sollen die Auswirkungen einiger Verteilungsannahmen auf die Parameter diskutiert werden. Bei festen, nicht zufälligen Variablenwerten betragen c_{22} , c_{23} und c_{24} 0. Daher gilt für die Parameter des Satzes 2.2.5: $\text{var}(T_2) = 0$, $\sigma_2^2 = \sigma_{21}^2$ und $\sigma_u^2 = \sigma_{u1}^2$.

Falls die $\bar{X}_{ij}$ stochastisch unabhängig sind, gilt $c_{23} = c_{24} = 0$. Bei Meßfehlermodellen (vgl. Abschnitt 3.2) kann man oft zusätzlich gleiche Varianzen für die $\bar{X}_{ij}$ annehmen; mit $v_2 := \text{var}(\bar{X}_{ij})$ gilt dann $c_{22} = \frac{n(n-1)}{2} v_2$. Wenn man weiterhin die $\bar{X}_{ij}$ als identisch verteilt ansieht, hat man den Standardfall von $\frac{K(K-1)}{2}$ unabhängigen Beobachtungen von Variablen mit derselben Wahrscheinlichkeitsverteilung, wobei allerdings die Zahl der Beobachtungen variieren kann. Bei Schätzung des Erwartungswertes $\mu_2 = E(\bar{X}_{ij}) = \frac{2}{n(n-1)} \theta_2$ durch $\mu_2^* = \frac{2}{n(n-1)} \theta_2^*$ muß also Satz 2.2.5 zu dem bekannten Ergebnis führen:

$$(2.2.47) \quad E(\mu_2^* - \mu_2)^2 = v_2 E\left(\frac{2}{K(K-1)}\right) ,$$

was man durch Einsetzen von $\sigma_2^2 = \frac{(n+1)(n-2)}{n(n-1)} v_2$ und $\sigma_u^2 = \frac{(n-1)(n-2)}{n} v_2$ zeigen kann.

2.3 Zulässigkeit, vollständige Klassen von Schätz- funktionen

Wie in Abschnitt 1.4 sollen zulässige Schätzfunktionen und eine kleine vollständige Klasse von Schätzfunktionen bestimmt werden, um anschließend insbesondere die Schätzfunktion θ_2^* und die Horvitz-Thompson-Schätzfunktion $\tilde{\theta}_2$ zu vergleichen. Wegen der Analogie zu Abschnitt 1.4 werden Beweise gelegentlich gekürzt und bei den Minimierungsproblemen werden die genauen Bezüge zu den benützten Büchern Ioffe, Tichomirov (1979) und Luenberger (1973) weggelassen; der Leser findet sie an den entsprechenden Stellen des Abschnittes 1.4.

Betrachtet wird nur die Klasse Ψ_2 der als Funktion der Zufallsvariablen $\bar{X}_{ij} S_i S_j$ linearen, symmetrischen Schätzfunktionen (vgl. Voraussetzung 2.1.2), die durch die Größen $\gamma_k = \gamma(k)$ mit $\alpha_k > 0$ vollständig bestimmt sind. Schätzfunktionen, die sich durch γ_k mit $\alpha_k = 0$ unterscheiden, werden wieder als eine Schätzfunktion betrachtet, und mit γ bezeichnen wir den $\ell \leq n$ dimensionalen Vektor mit den Komponenten γ_k für $\alpha_k > 0$. Es gelten die Voraussetzungen 2.1.1 - 2.1.3 und zugrunde gelegt wird die Gaußsche Verlustfunktion, falls nicht explizit eine andere Verlustfunktion genannt wird.

Nach Satz 2.2.5 hängt die Risikofunktion von der Verteilung W der Werte $(X_{12}, \dots, X_{ij}, \dots, X_{(n-1)n})$ nur über die Parameter σ_2^2 , σ_u^2 , $\text{var}(T_2)$ und θ_2^2 ab; Familien $\mathcal{W}$ von Verteilungen können wir also durch Parametermengen $\Lambda_2 \subset \mathbb{R}_+^4$ mit den Elementen $(\sigma_2^2, \sigma_u^2, \text{var}(T_2), \theta_2^2)$ beschreiben, wobei wegen

(2.2.23) stets $\sigma_u^2 \leq (n-1)^2 \sigma_2^2$ gelten muß und deshalb nicht jeder Wert aus $\mathbb{R}_+^4$ als Parameter in Frage kommt. Wir werden wieder voraussetzen, daß Λ_2 Elemente mit $\theta_2^2 \neq 0$ enthält. Die Risikofunktion läßt sich wie folgt darstellen:

$$(2.3.1) \quad R(\hat{\theta}_2, W) = R(\gamma, \sigma_2^2, \sigma_u^2, \text{var}(T_2), \theta_2^2) \\ = \sigma_2^2 f_1(\gamma) + \sigma_u^2 f_2(\gamma) + \text{var}(T_2) f_3(\gamma) + \theta_2^2 f_4(\gamma)$$

mit

$$(2.3.2) \quad f_1(\gamma) = \sum_k \alpha_k \gamma_k^2 \frac{k(k-1)(n-k)(n-k-1)}{2(n-2)(n-3)} \geq 0$$

$$(2.3.3) \quad f_2(\gamma) = \sum_k \alpha_k \gamma_k^2 \frac{k(k-1)(k-2)(n-k)}{(n-1)(n-2)(n-3)} \geq 0$$

$$(2.3.4) \quad f_3(\gamma) = \sum_k \alpha_k \gamma_k^2 \frac{k^2(k-1)^2}{n^2(n-1)^2} \geq 0$$

$$(2.3.5) \quad f_4(\gamma) = \sum_k \alpha_k \left(\gamma_k \frac{k(k-1)}{n(n-1)} - 1 \right)^2 \\ = \sum_k \alpha_k \left(\gamma_k^2 \frac{k^2(k-1)^2}{n^2(n-1)^2} - 2\gamma_k \frac{k(k-1)}{n(n-1)} + 1 \right) \geq 0,$$

wobei jeweils über alle k mit $\alpha_k > 0$ summiert wird. Wegen $\alpha_0 = \alpha_1 = 0$ gilt $f_3(\gamma) = 0$ nur, wenn alle γ_k mit $\alpha_k > 0$ Null betragen.

Hilfssatz 2.3.1

Die Funktionen f_1, f_2, f_3 und f_4 sind konvexe Funktionen von $\mathbb{R}^l$ in $\mathbb{R}$. Damit sind auch Linearkombinationen $c_1 f_1 + c_2 f_2 + c_3 f_3 + c_4 f_4 + c_5$ mit $c_1, c_2, c_3, c_4, c_5 \in \mathbb{R}$ und $c_1, c_2, c_3, c_4 \geq 0$ konvex und insbesondere ist die Risikofunktion für feste Parameter als Funktion von γ konvex.

Beweis

Die ersten und zweiten partiellen Ableitungen lauten:

$$(2.3.6) \quad \frac{\partial f_1}{\partial \gamma_k} = 2\alpha_k \gamma_k \frac{k(k-1)(n-k)(n-k-1)}{2(n-2)(n-3)}$$

$$(2.3.7) \quad \frac{\partial f_2}{\partial \gamma_k} = 2\alpha_k \gamma_k \frac{k(k-1)(k-2)(n-k)}{(n-1)(n-2)(n-3)}$$

$$(2.3.8) \quad \frac{\partial f_3}{\partial \gamma_k} = 2\alpha_k \gamma_k \frac{k^2(k-1)^2}{n^2(n-1)^2}$$

$$(2.3.9) \quad \frac{\partial f_4}{\partial \gamma_k} = 2\alpha_k \gamma_k \frac{k^2(k-1)^2}{n^2(n-1)^2} - 2\alpha_k \frac{k(k-1)}{n(n-1)}$$

$$(2.3.10) \quad \frac{\partial^2 f_i}{\partial \gamma_k \partial \gamma_j} = 0 \quad \text{für } i = 1, 2, 3, 4; k \neq j$$

$$(2.3.11) \quad \frac{\partial^2 f_1}{\partial \gamma_k \partial \gamma_k} = 2\alpha_k \frac{k(k-1)(n-k)(n-k-1)}{2(n-2)(n-3)} \geq 0$$

$$(2.3.12) \quad \frac{\partial^2 f_2}{\partial \gamma_k \partial \gamma_k} = 2\alpha_k \frac{k(k-1)(k-2)(n-k)}{(n-1)(n-2)(n-3)} \geq 0$$

$$(2.3.13) \quad \frac{\partial^2 f_3}{\partial \gamma_k \partial \gamma_k} = 2\alpha_k \frac{k^2(k-1)^2}{n^2(n-1)^2} > 0$$

$$(2.3.14) \quad \frac{\partial^2 f_4}{\partial \gamma_k \partial \gamma_k} = 2\alpha_k \frac{k^2(k-1)^2}{n^2(n-1)^2} > 0$$

Aus Gl. (2.3.10) - (2.3.14) folgt, daß die Matrizen der zweiten partiellen Ableitungen positiv semidefinit sind und somit gilt die Aussage des Satzes.

Im weiteren werden wir häufig Schätzfunktionen betrachten, bei denen γ_k für alle oder einige k folgende Gleichung erfüllt:

$$(2.3.15) \quad \gamma_k = \frac{1}{c(n-k)(n-k-1) + d(k-2)(n-k) + ek(k-1)}$$

mit $c \geq 0$, $d \geq 0$ und $e \geq \frac{1}{n(n-1)}$, z.B. werden Schätzfunktionen betrachtet, bei denen γ_n , γ_{n-1} oder eventuell γ_2 Gl.

(2.3.15) genügen und sonst $\gamma_k = 0$ gilt. Diese Aussagen sind nur sinnvoll, falls die Stichprobengröße K die extremen Werte n , $n-1$, 2 mit positiver Wahrscheinlichkeit annehmen kann. Da bei Stichprobenverfahren K diese Werte normalerweise nicht oder nur mit verschwindend kleinen Wahrscheinlichkeiten annimmt, haben diese Betrachtungen wenig praktischen Wert; aus mathematischen Gründen sind sie allerdings notwendig. Außerdem wird man kaum Schätzfunktionen mit $\gamma_k = 0$ für $\alpha_k > 0$ wählen, da sie für Stichproben vom Umfang k unabhängig von den beobachteten Werten als einzigen Schätzwert 0 annehmen. Speziell im nächsten Satz sind die wichtigen Aussagen im Teil (i) und (ii) enthalten, der Rest ist eine Betrachtung von bestimmten Ausnahmefällen.

Satz 2.3.2

- (i) Schätzfunktionen, die durch Gl. (2.3.15) für $k=2, \dots, n$ mit $c \geq 0$, $e \geq \frac{1}{n(n-1)}$ und $0 \leq d \leq 2c(n-1)$ definiert werden, sind gleichmäßig beste Schätzfunktionen bzgl. dem Parameterraum

$$(2.3.16) \quad \Lambda_{c,d,e} := \{ (\sigma_2^2, \sigma_u^2, \text{var}(T_2), \theta_2^2) : \theta_2^2 > 0, \quad$$

$$\text{var}(T_2) = \theta_2^2 (en(n-1)-1), \sigma_u^2 = \theta_2^2 d \frac{(n-2)(n-3)}{n},$$

$$\sigma_2^2 = \theta_2^2 c \frac{2(n-2)(n-3)}{n(n-1)} \}.$$

Falls $\Lambda_2 \cap \Lambda_{c,d,e} \neq \emptyset$, sind sie zulässig.

(ii) Falls Λ_2 Parameter $(\sigma_2^2, \sigma_u^2, \text{var}(T_2), 0)$ mit $\text{var}(T_2) > 0$ enthält, ist die durch $\gamma_k = 0, k=2, \dots, n$ definierte Schätzfunktion bzgl. Λ_2 zulässig.

(iii) Falls Λ_2 Parameter $(\sigma_2^2, 0, 0, 0)$ mit $\sigma_2^2 > 0$ enthält, sind die Schätzfunktionen zulässig, die durch $\gamma_k = 0$ für $k < n-1$,

$$(2.3.17) \quad \gamma_{n-1} = \frac{1}{\frac{\sigma_u^2}{\theta_2^2} \frac{n}{n-2} + \frac{\text{var}(T_2) + \theta_2^2}{\theta_2^2} \frac{n-2}{n}} \quad \text{und}$$

$$(2.3.18) \quad \gamma_n = \frac{\theta_2^2}{\text{var}(T_2) + \theta_2^2}$$

für $(\sigma_2^2, \sigma_u^2, \text{var}(T_2), \sigma_2^2) \in \Lambda_2$ mit $\theta_2 > 0$ definiert werden.

(iv) Falls Λ_2 Parameter $(\sigma_2^2, \sigma_u^2, 0, 0)$ mit $\sigma_2^2 > 0$ und $\sigma_u^2 > 0$ enthält, sind die Schätzfunktionen zulässig, die durch $\gamma_k = 0$ für $k < n$, durch γ_n gemäß Gl. (2.3.18) definiert sind.

Beweis

Für feste Parameter $(\sigma_2^2, \sigma_u^2, \text{var}(T_2), \theta_2^2)$ mit $\theta_2^2 > 0$ oder $\text{var}(T_2) > 0$ gilt:

$$(2.3.19) \quad \begin{aligned} \frac{\partial R}{\partial \gamma_k} = & 2\alpha_k \gamma_k \left[\sigma_2^2 \frac{k(k-1)(n-k)(n-k-1)}{2(n-2)(n-3)} \right. \\ & + \sigma_u^2 \frac{k(k-1)(k-2)(n-k)}{(n-1)(n-2)(n-3)} \\ & \left. + (\text{var}(T_2) + \theta_2^2) \left(\frac{k(k-1)}{n(n-1)} \right)^2 \right] - 2\alpha_k \theta_2^2 \frac{k(k-1)}{n(n-1)}. \end{aligned}$$

Die einzige Nullstelle der ersten Ableitung lautet:

$$(2.3.20) \quad \gamma_k = \theta_2^2 \left[\sigma_2^2 \frac{n(n-1)(n-k)(n-k-1)}{2(n-2)(n-3)} + \sigma_u^2 \frac{n(k-2)(n-k)}{(n-2)(n-3)} \right. \\ \left. (\text{var}(T_2) + \theta_2^2 \frac{k(k-1)}{n(n-1)}) \right]^{-1} \quad \text{für } k=2, \dots, n.$$

Aus Gl. (2.3.10) - (2.3.14) folgt, daß die Matrix der zweiten partiellen Ableitungen positiv definit ist und somit gibt Gl. (2.3.20) das einzige globale Minimum an.

zu (i): Für $\theta_2^2 > 0$ kann man (2.3.20) auch in der Form

$$(2.3.15) \text{ mit } c = \frac{n(n-1)}{2(n-2)(n-3)} \frac{\sigma_2^2}{\theta_2^2}, \quad d = \frac{n}{(n-2)(n-3)} \frac{\sigma_u^2}{\theta_2^2}$$

$$\text{und } e = \frac{1}{n(n-1)} \frac{\text{var}(T_2) + \theta_2^2}{\theta_2^2} \text{ schreiben. Die so}$$

definierte Schätzfunktion ist für

$(\sigma_2^2, \sigma_u^2, \text{var}(T_2), \theta_2^2) \in \Lambda_{c,d,e}$ die einzige gleichmäßig beste Schätzfunktion und daher bezüglich Λ_2 zulässig, falls $\Lambda_2 \cap \Lambda_{c,d,e} \neq \emptyset$.

zu (ii): Für $\text{var}(T_2) > 0$ und $\theta_2^2 = 0$ folgt aus (2.3.20)

$$\gamma_k = 0, \quad k=2, \dots, n.$$

zu (iii): Es gelte $\alpha_n > 0$ und $\alpha_{n-1} > 0$; $\tilde{\gamma} \in \mathbb{R}^l$ sei ein beliebiger Vektor und $\gamma \in \mathbb{R}^l$ sei ein Vektor, der die Bedingungen des Teils (iii) erfüllt. Für $\sigma_2^2 > 0$ gilt $0 = R(\gamma, \sigma_2^2, 0, 0, 0) \geq R(\tilde{\gamma}, \sigma_2^2, 0, 0, 0)$ genau dann, wenn $\tilde{\gamma}_k = 0$ für $k < n-1$. Für $\tilde{\gamma} \in \mathbb{R}^l$, die dieser Bedingung genügen, gilt:

$$\begin{aligned}
 (2.3.21) \quad R(\tilde{\gamma}, \sigma_2^2, \sigma_u^2, \text{var}(T_2), \theta_2^2) \\
 = \alpha_{n-1} \left[\sigma_u^2 \tilde{\gamma}_{n-1}^2 + \text{var}(T_2) \tilde{\gamma}_{n-1}^2 \left(\frac{n-2}{n} \right)^2 \right. \\
 \left. + \theta_2^2 \left(\tilde{\gamma}_{n-1}^2 \left(\frac{n-2}{n} \right)^2 - 2 \tilde{\gamma}_{n-1} \frac{n-2}{n} + 1 \right) \right] \\
 + \alpha_n \left[\text{var}(T_2) \tilde{\gamma}_n^2 + \theta_2^2 (\tilde{\gamma}_n - 1)^2 \right].
 \end{aligned}$$

Wie in Teil (i) kann man zeigen, daß für feste Parameter mit $\theta_2^2 > 0$ das einzige globale Minimum dieser Funktion (von $(\tilde{\gamma}_{n-1}, \tilde{\gamma}_n)$) durch Gleichung (2.3.17) und (2.3.18) definiert wird. Alle Schätzfunktionen, die nicht den Bedingungen des Teils (iii) genügen, sind also entweder für Parameter $(\sigma_2^2, 0, 0, 0)$ oder $(\sigma_2^2, \sigma_u^2, \text{var}(T_2), \theta_2^2)$ schlechter als die durch γ definierten Schätzfunktionen.

zu (iv): Analog zum Beweis von Teil (iii) kann man zeigen, daß Schätzfunktionen, die nicht den Bedingungen des Teils (iv) genügen, entweder für $(\sigma_2^2, \sigma_u^2, 0, 0) \in \Lambda_2$ mit $\sigma_2^2 > 0$, $\sigma_u^2 > 0$ oder für Parameter $(\sigma_2^2, \sigma_u^2, \text{var}(T_2), \theta_2^2) \in \Lambda_2$ schlechter als die angegebenen sind.

Bemerkung 2.3.3

Die Gleichungen (2.3.17) und (2.3.18) sind Spezialfälle von (2.3.15) für $k=n-1$ und $k=n$ mit c beliebig,

$$d = \frac{\sigma_u^2}{\theta_2^2} \frac{n}{(n-2)(n-3)} \quad \text{und} \quad e = \frac{\text{var}(T_2) + \theta_2^2}{\theta_2^2} \frac{1}{n(n-1)}.$$

(2.2.23) nicht $\sigma_2^2 = 0$ und $\sigma_u^2 > 0$ auftreten kann, brauchte

dieser Fall nicht untersucht zu werden. Die Festlegung von $\gamma_n=1$ ist für feste Werte (d.h. $\text{var}(T_2)=0$) wieder intuitiv klar, falls die Werte aller Paare von Untersuchungseinheiten beobachtet werden, ist ihre Summe gleich dem wahren θ_2 und man hat keinen Schätzfehler.

Bemerkung 2.3.4

Falls bei $\theta_2^2 > 0$ die Variationskoeffizienten

$$v_2^2 = \frac{\sigma_2^2}{\theta_2^2}, \quad u^2 = \frac{\sigma_u^2}{\theta_2^2} \text{ bei festen Werten oder allgemeiner}$$

$$v_2^2, u^2 \text{ und } w_2^2 = \frac{\text{var}(T_2)}{\theta_2^2} \text{ bekannt sind, sind durch}$$

(2.3.22)

$$\gamma_k = \frac{1}{v_2^2 \frac{n(n-1)(n-k)(n-k-1)}{2(n-2)(n-3)} + u^2 \frac{n(k-2)(n-k)}{(n-2)(n-3)} + (w_2^2 + 1) \frac{k(k-1)}{n(n-1)}}$$

$k=2, \dots, n$ gleichmäßig beste Schätzfunktionen definiert.

Satz 2.3.5

Die Menge der linearen, symmetrischen Schätzfunktionen mit:

$$(2.3.23) \quad 0 \leq \gamma_k \leq \frac{n(n-1)}{k(k-1)} \quad \text{für } k = 2, \dots, n$$

ist vollständig.

Beweis

Analog zum Beweis von Satz 1.4.5 läßt sich zu einer durch $\tilde{\gamma}$ definierten Schätzfunktion, die nicht zur angegebenen

vollständigen Menge gehört, eine gleichmäßig bessere Schätzfunktion konstruieren, indem man $\gamma_k = |\tilde{\gamma}_k|$ für $|\tilde{\gamma}_n| \leq \frac{n(n-1)}{k(k-1)}$ und $\gamma_k = \frac{n(n-1)}{k(k-1)}$ für $|\tilde{\gamma}_k| > \frac{n(n-1)}{k(k-1)}$ setzt.

Bemerkung 2.3.6

Falls $\text{var}(K) > 0$, gilt $\frac{n(n-1)}{E(K(K-1))} > \frac{n(n-1)}{k(k-1)}$ für $k(k-1) > E(K(K-1))$ und deshalb existiert nach Satz 2.3.5 eine gleichmäßig bessere als die durch $\gamma_k = \frac{n(n-1)}{E(K(K-1))}$ definierte Horvitz-Thompson-Schätzfunktion.

Satz 2.3.7

(i) Es sei:

$$(2.3.24) \quad \Psi'_2 := \{ \hat{\theta}_2 = \gamma(K) \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \bar{x}_{ij} S_i S_j : \}$$

γ_k genügen Gl. (2.3.15) für $k = 2, \dots, n$

oder γ_n und γ_{n-1} genügen Gl. (2.3.15)

und $\gamma_k = 0$ sonst

oder γ_2 und γ_n genügen Gl. (2.3.15)

und $\gamma_k = 0$ sonst

oder $\gamma_n \in [0, 1]$ und $\gamma_k = 0$ sonst}.

Ψ'_2 ist vollständig.

(ii) Es sei:

$$(2.3.25) \quad \Psi''_2 := \{ \hat{\theta}_2 \in \Psi'_2 : \gamma_k \text{ genügen Gl. (2.3.15)} \\ \text{mit } e = \frac{1}{n(n-1)} \}$$

oder $\gamma_n = 1$ und $\gamma_k = 0$ sonst}.

Für Parameterräume $\Lambda_2 \subset \{(\sigma_2^2, \sigma_u^2, \text{var}(T_2), \theta_2^2) :$

$$\sigma_2^2 \geq 0, \theta_2^2 \geq 0, 0 \leq \sigma_u^2 \leq (n-1)^2 \sigma_2^2, \text{var}(T_2) = 0\}$$

ist Ψ_2'' vollständig.

Beweis

zu (i): Bei der Betrachtung können wir ohne Einschränkung der Allgemeinheit Gl. (2.3.23) voraussetzen. Für einen Vektor γ gilt:

$$f_1(\gamma) = 0 \text{ genau dann, wenn } \gamma_k = 0 \text{ für } k < n-1$$

$$f_2(\gamma) = 0 \text{ genau dann, wenn } \gamma_k = 0 \text{ für } 2 < k < n$$

$$f_3(\gamma) = 0 \text{ genau dann, wenn } \gamma_k = 0 \text{ für } k=2, \dots, n$$

Für Schätzfunktionen nicht aus Ψ_2' ist also stets $f_1(\gamma) > 0$ oder $f_2(\gamma) > 0$; wir unterscheiden die drei möglichen Fälle:

Fall 1:

Durch $\tilde{\gamma}$ sei eine Schätzfunktion $\tilde{\theta}_2 \notin \Psi_2'$ definiert mit $f_1(\tilde{\gamma}) > 0$, $f_2(\tilde{\gamma}) > 0$ und $f_3(\tilde{\gamma}) > 0$. Wir konstruieren eine bessere Schätzfunktion, indem wir f_4 unter den Nebenbedingungen

$$(2.3.26) \quad f_1(\gamma) - f_1(\tilde{\gamma}) \leq 0,$$

$$(2.3.27) \quad f_2(\gamma) - f_2(\tilde{\gamma}) \leq 0 \text{ und}$$

$$(2.3.28) \quad f_3(\gamma) - f_3(\tilde{\gamma}) \leq 0 \text{ minimieren.}$$

Nach dem Satz von Kuhn-Tucker haben wir für feste, unbekannte Lagrangemultiplikatoren $\lambda_1, \lambda_2, \lambda_3 \geq 0$ die Lagrangefunktion

$$(2.3.29) \quad L(\gamma, \lambda_1, \lambda_2, \lambda_3) = R(\gamma, \lambda_1, \lambda_2, \lambda_3, 1) - \lambda_1 f_1(\tilde{\gamma}) \\ - \lambda_2 f_2(\tilde{\gamma}) - \lambda_3 f_3(\tilde{\gamma})$$

zu minimieren. Im Beweis von Satz 2.3.2 ergab sich als einziges Minimum dieser Funktion

$$(2.3.30) \quad \gamma_k = \frac{1}{c(n-k)(n-k-1) + d(k-2)(n-k) + e k(k-1)}, k=2, \dots, n$$

mit $c = \frac{n(n-1)}{2(n-2)(n-3)} \lambda_1$, $d = \frac{n}{(n-2)(n-3)} \lambda_2$ und

$e = \frac{\lambda_3 + 1}{n(n-1)}$. Für γ_k gilt Gl. (2.3.23). Wie im Beweis von Satz 1.4.7 kann man folgern, daß die durch (2.3.30) definierte Schätzfunktion besser als die durch $\tilde{\gamma}$ definierte Schätzfunktion ist.

Fall 2:

Für $\tilde{\gamma}$ gelte $\tilde{\gamma}_k = 0$ für $k < n-1$ und $\tilde{\gamma}_{n-1} \neq 0$ (d.h. $f_1(\tilde{\gamma}) = 0$ und $f_2(\tilde{\gamma}) > 0$), $\tilde{\gamma}_{(n-1)}$ und $\tilde{\gamma}_n$ erfüllen nicht Gl. (2.3.15). Wie bei Fall 1 kann man nun eine gleichmäßig bessere Schätzfunktion konstruieren mit $\gamma_k = 0$ für $k < n-1$ und γ_{n-1} und γ_n gemäß Gl. (2.3.15) bzw. (2.3.30). Wir betrachten dabei die Funktionen f_1, f_2, f_3, f_4 nur auf der konvexen Menge $\{\gamma \in \mathbb{R}^L : \gamma_k = 0 \text{ für } k < n-1\}$; die Nebenbedingung (2.3.26) fällt weg, da auf dieser Menge $f_1(\gamma) = 0$ stets gilt und bei der Lösung des Minimierungsproblems $(0, \dots, 0, \gamma_{n-1}, \gamma_n)$ gilt für γ_{n-1} und γ_n Gl. (2.3.15). Die Wahl von c in Gl. (2.3.15) ist für $k=n$ oder $k=n-1$ beliebig, da in diesen Fällen $n-k$ oder $n-k-1$ gleich 0 wird.

Fall 3:

Für $\tilde{\gamma}$ gelte $\tilde{\gamma}_2 \neq 0$ und $\gamma_k = 0$ für $2 < k < n$ (d.h. $f_1(\tilde{\gamma}) > 0$, $f_2(\tilde{\gamma}) = 0$). $\tilde{\gamma}_2$ und $\tilde{\gamma}_n$ erfüllen nicht Gl. (2.3.15). Wir betrachten die Funktionen f_1, f_2, f_3, f_4 nur auf der konvexen Menge $\{\gamma \in \mathbb{R}^l : \gamma_k = 0 \text{ für } 2 < k < n\}$; die Nebenbedingung (2.3.27) fällt weg und analog zu Fall 1 erhalten wir als Lösung des Minimierungsproblems $(\gamma_2, 0, \dots, 0, \gamma_n)$, wobei γ_2 und γ_n gemäß Gl. (2.3.15) definiert sind. Die Wahl von d ist in Gl. (2.3.15) für $k=2$ oder $k=n$ beliebig, da dann entweder $k-2$ oder $n-k$ gleich 0 wird.

zu (ii): Es sei $\Psi_2''' := \{\gamma \in \mathbb{R}^l : \gamma_k = 0 \text{ für } k < n\}$. Da stets $\text{var}(T_2) = 0$ gilt, ist die durch $\gamma_k = 0$ für $k < n$ und $\gamma_n = 1$ definierte Schätzfunktion die gleichmäßig beste bezüglich Ψ_2''' . Zu Schätzfunktionen $\tilde{\theta}_2 \notin \Psi_2'' \cup \Psi_2'''$ läßt sich analog zu Teil (i) eine gleichmäßig bessere Schätzfunktion konstruieren, wobei allerdings bei dem Minimierungsproblem wegen $\text{var}(T_2) = 0$ Nebenbedingung (2.3.28) und dadurch der Lagrange Multiplikator λ_3 wegfällt (bzw. $\lambda_3 = 0$) und somit $e = \frac{1}{n(n-1)}$ gesetzt werden kann.

Bemerkung 2.3.8

Bei der Wahl einer Schätzfunktion kann man sich also auf Schätzfunktionen aus Ψ_2' bzw. Ψ_2'' beschränken. Allerdings sind Ψ_2' und Ψ_2'' stets echt größer als die in Satz 2.3.2 bestimmte Menge von zulässigen Schätzfunktionen: Ψ_2' und Ψ_2'' enthalten Schätzfunktionen, für die Gl. (2.3.15) für alle k mit $d > 2c(n-1)$ gilt und Schätzfunktionen mit $\gamma_k = 0$ für $2 < k < n$ und $\gamma_2 \neq 0$. Für Parameterräume $\Lambda_2 \neq \Lambda_{c,d,e}$ haben wir also in diesem Abschnitt weder die Menge aller

zulässigen, noch eine minimale vollständige Klasse von Schätzfunktionen bestimmt. Wie schon gesagt wurde, sind die Schätzfunktionen mit $\gamma_k=0$ für $\alpha_k>0$ praktisch nicht relevant, man wird also Schätzfunktionen wählen, für die Gl. (2.3.15) für alle k gilt und sich bei diesen Schätzfunktionen auch auf die zulässigen beschränken. Durch Wahl der Größen c, d, e kann man die Gewichtungen der Parameter $(\sigma_2^2, \sigma_u^2, \text{var}(T_2), \theta_2^2)$ wieder so bestimmen, daß bestimmte Parameter wenig Einfluß auf die Risikofunktion haben.

Bemerkung 2.3.9

In der Praxis wird oft der Stichprobenumfang fest vorgegeben und als symmetrisches Stichprobenverfahren kommt dann nur noch das k -malige Ziehen ohne Zurücklegen in Frage. Die in den Sätzen 2.3.5 und 2.3.7 bestimmten vollständigen Klassen sind in diesem Fall gleich; Satz 2.3.7 bringt also keinen Informationsgewinn. Außerdem werden bei fest vorgegebenem Stichprobenumfang die Bemerkung über die Unzulässigkeit der Horvitz-Thompson-Schätzfunktion und wegen $\theta_2^* = \tilde{\theta}_2$ der noch folgende Vergleich beider Schätzfunktionen gegenstandslos.

Falls der Stichprobenumfang K nur 2 oder 3 verschiedene Werte annehmen kann, ist es möglich, bei bekanntem γ_k mit $0 < \gamma_k \leq \frac{n(n-1)}{k(k-1)}$ für k mit $\alpha_k>0$ die Gleichungen der Form (2.3.15) nach c, d und e aufzulösen. Wie man z.B. für $\gamma_{k_1} = \frac{n(n-1)}{k_1(k_1-1)}, \gamma_{k_2} = \frac{n(n-1)}{k_2(k_2-1)}, \gamma_k=0$ sonst nachprüfen kann, gilt für die Lösungen allerdings nicht stets $c \geq 0, d \geq 0$ und $e \geq \frac{1}{n(n-1)}$; ψ_2'' ist also kleiner als die in Satz 2.3.5 angegebene vollständige Menge.

Bemerkung 2.3.10

Es gibt nur eine erwartungstreue Schätzfunktion in Ψ'_2 bzw. Ψ''_2 : θ_2^* mit $\gamma^*(K) = \frac{n(n-1)}{k(k-1)}$ (d.h. $c=d=0$, $e = \frac{1}{n(n-1)}$ in Gl. (2.3.15)). Für alle anderen Schätzfunktionen $\hat{\theta}_2 \in \Psi'_2$ gilt $(0 \leq) \gamma_k \leq \gamma_k^*$ mit Ungleichheit für mindestens ein k mit $\alpha_k > 0$ und somit gilt für $\hat{\theta}_2 \in \Psi'_2$, $\hat{\theta}_2 \neq \theta_2^*$:

$$(2.3.31) \quad |E(\hat{\theta}_2)| = |\theta_2| \sum \alpha_k \gamma_k \frac{k(k-1)}{n(n-1)} < |\theta_2| \sum \alpha_k \gamma_k^* \frac{k(k-1)}{n(n-1)} \leq |\theta_2|,$$

die Schätzfunktionen verzerren also nach 0 hin.

Bemerkung 2.3.11

Wir kommen nun zum Vergleich von $\tilde{\theta}_2$ und θ_2^* für Stichprobenverfahren mit $\text{var}(K) > 0$. Beide Schätzfunktionen sind erwartungstreu, für die bedingten Erwartungswerte gilt allerdings:

$$(2.3.32) \quad E(\theta_2^* | K) = \theta_2 \quad \text{und}$$

$$(2.3.33) \quad E(\tilde{\theta}_2 | K=k) \neq \theta_2 \quad \text{für } k \neq E(K).$$

Nach Kenntnis des Stichprobenumfangs muß man für $k \neq E(K)$ bei der Horvitz-Thompson-Schätzfunktion mit einer systematischen Verzerrung rechnen. Für θ_2^* gilt $f_4(\gamma^*) = 0$, die Werte der Risikofunktion hängen also im Gegensatz zu $\tilde{\theta}_2$ nicht vom wahren θ_2 ab. Ein direkter Vergleich der Varianzen von $\tilde{\theta}_2$ und θ_2^* ist hier nicht durchgeführt. Es ist zu vermuten, daß ähnlich wie bei $\tilde{\theta}_1$ und θ_1^* (vgl. Bem. 1.4.10) die Streuungen σ_2^2 und σ_u^2 bei $\tilde{\theta}_2$ geringer gewichtet werden und deshalb für

bestimmte Parameter $\tilde{\theta}_2$ besser als θ_2^* ist. Da $\tilde{\theta}_2$ nicht in den in Satz 2.3.7 bestimmten vollständigen Klassen liegt, gibt es allerdings eine nicht erwartungstreue Schätzfunktion aus Ψ_2' bzw. Ψ_2'' , die gleichmäßig besser als $\tilde{\theta}_2$ ist; es gibt also wenig triftige Gründe, sich für die Horvitz-Thompson-Schätzfunktion zu entscheiden. Da außer den hier genannten Gründen θ_2^* eine einfache und gut interpretierbare Form hat, ist meistens die Wahl der zulässigen Schätzfunktion θ_2^* naheliegend.

Wie in Satz 1.4.11 ist es wieder möglich zu zeigen, daß θ_2^* für die logarithmische Verlustfunktion eine gleichmäßig beste Schätzfunktion ist.

Satz 2.3.12

Bei Zugrundelegung der logarithmischen Verlustfunktion (Gl. (1.2.6)) ist θ_2^* mit $\gamma_k^* = \frac{n(n-1)}{k(k-1)}$, $k=1, \dots, n$ eine gleichmäßig beste Schätzfunktion, wobei $\theta_2 > 0$ und $P(\hat{\theta}_2 > 0) = 1$ vorausgesetzt werden muß.

Beweis

Bei logarithmischer Verlustfunktion mit $g(\theta) \equiv 1$ lautet die Risikofunktion:

$$\begin{aligned}
 (2.3.34) \quad R(\hat{\theta}_2, \sigma_2^2, \sigma_u^2, \text{var}(T_2), \theta_2^2) \\
 &= E\left(\frac{\hat{\theta}_2}{\theta_2}\right) - E\left(\log\left(\frac{\hat{\theta}_2}{\theta_2}\right)\right) - 1 \\
 &= E\left(\gamma(K) \frac{K(K-1)}{n(n-1)}\right) - E\left(\log\left(\gamma(K) \frac{K(K-1)}{n(n-1)}\right)\right) - 1 \geq 0
 \end{aligned}$$

Da $\gamma_k > 0$ für k mit $\alpha_k > 0$ stets gilt, nimmt sie ihr Minimum unabhängig von $\sigma_2^2, \sigma_u^2, \text{var}(T_2)$ und θ_2^2 für $\gamma^*(K) = \frac{n(n-1)}{K(K-1)}$ an.

2.4 Die Schätzung anderer Parameter

Auch bei der Schätzung der anderen in der Risikofunktion vorkommenden Parameter σ_2^2 , σ_u^2 , $\text{var}(T_2)$ und θ_2^2 gehen wir wie in Kapitel I (Abschnitt 1.5) vor, können deshalb die Argumentation etwas kürzen und uns auf die Angabe und Herleitung erwartungstreuer Schätzfunktionen beschränken.

Wir beginnen mit der Schätzung der Parameter a_{22} , a_{23} und a_{24} . Wegen der Unabhängigkeit der Werte vom Stichprobenverfahren und wegen Gl. (1.1.28) sind folgende Schätzfunktionen erwartungstreu:

$$(2.4.1) \quad a_{22}^* := \frac{n(n-1)}{K(K-1)} \sum_{i=1}^n \sum_{j>i}^n \bar{x}_{ij}^2 S_i S_j$$

$$(2.4.2) \quad a_{23}^* := \frac{n(n-1)(n-2)}{K(K-1)(K-2)} \sum_{i=1}^n \sum_{j>i}^n \sum_{\substack{\ell=1 \\ \ell \neq i \\ \ell \neq j}}^n \bar{x}_{ij} (\bar{x}_{i\ell} + \bar{x}_{j\ell}) S_i S_j S_\ell$$

$$(2.4.3) \quad a_{24}^* := \frac{n(n-1)(n-2)(n-3)}{K(K-1)(K-2)(K-3)} \sum_{i=1}^n \sum_{j>i}^n \sum_{\substack{\ell=1 \\ \ell \neq i \\ \ell \neq j}}^n \sum_{\substack{m>\ell \\ m \neq i \\ m \neq j}}^n \bar{x}_{ij} \bar{x}_{\ell m} S_i S_j S_\ell S_m$$

Nach Gl. (2.2.20) und (2.2.21) erhalten wir daraus erwartungstreue Schätzfunktionen für σ_2^2 und σ_u^2 :

$$(2.4.4) \quad \sigma_2^{2*} := \frac{2(n+1)(n-2)}{n^2(n-1)^2} a_{22}^* - \frac{4}{n^2(n-1)^2} (a_{23}^* + a_{24}^*)$$

$$(2.4.5) \quad \sigma_u^{2*} := \frac{2(n-2)}{n^2} a_{22}^* + \frac{n-4}{n^2} a_{23}^* - \frac{4}{n^2} a_{24}^*$$

In Gl. (1.5.5) hatten wir als Schätzfunktion σ_1^{2*} für die Streuung σ_1^2 die mit einer Korrektur versehene Varianz der beobachteten Werte in der Stichprobe erhalten; bei der Schätzung von σ_2^2 und σ_u^2 haben wir eine nicht so anschauliche Schätzfunktion erhalten und aus den Varianzen der beobachteten Werte der Stichprobe erhält man keine erwartungstreuen Schätzfunktionen für σ_2^2 und σ_u^2 , z.B. gilt für

$$(2.4.6) \quad \sigma_2^{2**} := \delta(K) \sum_{i=1}^n \sum_{j>i}^n \left(\bar{x}_{ij} - \frac{2}{K(K-1)} \sum_{\ell=1}^n \sum_{m>\ell}^n \bar{x}_{\ell m} s_{\ell} s_m \right)^2 s_i s_j$$

$$\begin{aligned} (2.4.7) \quad E(\sigma_2^{2**}) &= \sum_{i=1}^n \sum_{j>i}^n E(\bar{x}_{ij}^2) E(\delta(K) s_i s_j) \\ &- E \left[\sum_{i=1}^n \sum_{j>i}^n \sum_{\ell=1}^n \sum_{m>\ell}^n \bar{x}_{ij} \bar{x}_{\ell m} \frac{4\delta(K)}{K(K-1)} s_i s_j s_{\ell} s_m \right. \\ &\quad \left. - \sum_{i=1}^n \sum_{j>i}^n \sum_{\ell=1}^n \sum_{m>\ell}^n \bar{x}_{ij} \bar{x}_{\ell m} \frac{2\delta(K)}{K(K-1)} s_i s_j s_{\ell} s_m \right] \\ &= a_{22} E \left(\frac{\delta(K) (K+1) (K-2)}{n(n-1)} \right) \\ &- a_{23} E \left(\frac{2\delta(K) (K-2)}{n(n-1) (n-2)} \right) \\ &- a_{24} E \left(\frac{2\delta(K) (K-2) (K-3)}{n(n-1) (n-2) (n-3)} \right) \end{aligned}$$

Insbesondere weil a_{23} und a_{24} verschieden stark gewichtet sind, läßt sich keine Funktion $\delta(K)$ finden, so daß Gl. (2.2.20) erfüllt ist; σ_2^{2**} kann also im allgemeinen nicht durch geeignete Wahl von $\delta(K)$ zu einer erwartungstreuen Schätzfunktion gemacht werden.

Aus den gleichen Gründen wie in Abschnitt 1.5 haben wir ohne zusätzliches Wissen keine Möglichkeit, aus den beobachteten Werten der $\bar{X}_{ij}$ in der Stichprobe weitere Informationen über die Parameter $\text{var}(T_2)$, $\text{var}(\theta_2^*)$, $\sum_{i=1}^n \sum_{j>i}^n \text{var}(\bar{X}_{ij})$ usw. zu erhalten. Falls man $\text{var}(T_2)$ kennt oder erwartungstreu schätzen kann, kann man θ_2^2 und das Risiko einer Schätzfunktion schätzen; wir erhalten als Schätzfunktion $(\theta_2^2)^*$ für θ_2^2 und $\widehat{\text{var}}(\theta_2^*)$ für $\text{var}(\theta_2^*)$ (siehe Gl. (2.2.10) und Bem. 2.2.6)

$$(2.4.8) \quad (\theta_2^2)^* := a_{22}^* + a_{23}^* + a_{24}^* - \text{var}(T_2)$$

$$(2.4.9) \quad \widehat{\text{var}}(\theta_2^*) = \sigma_2^{2*} \frac{n^2 (n-1)^2}{2(n-2)(n-3)} E\left(\frac{(n-K)(n-K-1)}{K(K-1)}\right) \\ + \sigma_u^{2*} \frac{n^2 (n-1)}{(n-2)(n-3)} E\left(\frac{(K-2)(n-K)}{K(K-1)}\right) \\ + \text{var}(T_2)$$

wobei $\text{var}(T_2)$ gegebenenfalls durch eine erwartungstreue Schätzfunktion ersetzt wird.

Wenn die Variablenwerte in der Grundgesamtheit fest vorgegeben sind, weiß man, daß $\text{var}(T_2)=0$ gilt. Falls die $\bar{X}_{ij}$ oder U_i identisch verteilt und unabhängig voneinander sind, kann man mittels Gl. (2.2.44) oder (2.2.45) aus σ_2^{2*} bzw. σ_u^{2*} erwartungstreue Schätzfunktionen für $\text{var}(T_2)$ gewinnen, weil

$$\text{dann } \sigma_2^2 = \frac{2(n+1)(n-1)}{n^2(n-1)^2} \text{var}(T_2) \text{ bzw. } \sigma_u^2 = \frac{2(n-1)}{n^2} \text{var}(T_2) \text{ gilt.}$$

Es sei darauf hingewiesen, daß $\text{cov}(U_i, U_j) = \text{var}(\bar{X}_{ij})$ aus der Unabhängigkeit der $\bar{X}_{ij}$ folgt. Deshalb können beide obengenannten Voraussetzungen gleichzeitig nur zutreffen, wenn $\bar{X}_{ij} = x$ konstant für alle $i, j \in \Omega$ ist.

Wir wollen wieder den Fall untersuchen, daß zu jedem Paar von Untersuchungseinheiten in der Stichprobe unabhängig voneinander mehrere Realisierungen der Zufallsvariablen beobachtet werden. $\bar{X}_{ijv}, v=1, \dots, \kappa$ seien die zur v -ten Beobachtung des Wertes des Paares $\{i, j\}$ gehörigen Zufallsvariablen. Für jedes $\{i, j\}$ seien die Variablen $\bar{X}_{ijv}, v=1, \dots, \kappa$ identisch verteilt und die Mengen von Zufallsvariablen $\{\bar{X}_{ijv} : i \in \Omega, j \in \Omega, j > i\} \quad v=1, \dots, \kappa$ seien unabhängig voneinander. Wir definieren:

$$(2.4.10) \quad \bar{X}_{ij} = \frac{1}{\kappa} \sum_{v=1}^{\kappa} \bar{X}_{ijv}$$

und für $\{i, j\} \in \mathcal{CS}$ werden alle $\bar{X}_{ijv}, v=1, \dots, \kappa$ und damit auch $\bar{X}_{ij}$ beobachtet. Die bisherigen Ergebnisse liefern Schätzfunktionen für $E\left(\sum_{i=1}^n \sum_{j>i}^n \frac{1}{\kappa} \sum_{v=1}^{\kappa} \bar{X}_{ijv}\right) = E\left(\sum_{i=1}^n \sum_{j>i}^n \bar{X}_{ij}\right) = \theta_2$, für

die Streuung σ_2^2 der $\bar{X}_{ij}$ und für die Streuung σ_u^2 von

$$U_i = \sum_{\substack{j=1 \\ j \neq i}}^n \bar{X}_{ij}.$$

Die Varianzen der $\bar{X}_{ij}$ können wir erwartungs-

treu durch

$$(2.4.11) \quad \sigma_{ij}^{2*} := \frac{1}{\kappa(\kappa-1)} \sum_{v=1}^{\kappa} (\bar{X}_{ijv} - \bar{X}_{ij})^2,$$

die Kovarianzen von X_{ij} und $X_{\ell m}$ durch

$$(2.4.12) \quad \sigma_{ij, \ell m}^* := \frac{1}{\kappa(\kappa-1)} \sum_{v=1}^{\kappa} (\bar{X}_{ijv} - \bar{X}_{ij})(\bar{X}_{\ell mv} - \bar{X}_{\ell m})$$

schätzen, falls $\{i, j\} \in \mathcal{CS}$ bzw. $\{i, j, \ell, m\} \in \mathcal{CS}$ (vgl. Stenger (1971) S. 31 u. 33).

Als erwartungstreue Schätzfunktionen C_{22}^* für $\sum_{i=1}^n \sum_{j>i}^n \text{var}(\bar{X}_{ij})$,

C_{23}^* für $\sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \sum_{\substack{\ell=1 \\ \ell \neq i \\ \ell \neq j}}^n \text{cov}(\bar{X}_{ij}, \bar{X}_{i\ell})$ und

C_{24}^* für $\sum_{i=1}^n \sum_{j>i}^n \sum_{\substack{\ell=1 \\ \ell \neq i \\ \ell \neq j}}^n \sum_{\substack{m>\ell \\ m \neq i \\ m \neq j}}^n \text{cov}(X_{ij}, X_{\ell m})$ erhalten wir dann:

$$(2.4.13) \quad C_{22}^* = \frac{n(n-1)}{K(K-1)} \sum_{i=1}^n \sum_{j>i}^n \sigma_{ij}^* S_i S_j$$

$$(2.4.14) \quad C_{23}^* = \frac{n(n-1)(n-2)}{K(K-1)(K-2)} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \sum_{\substack{\ell=1 \\ \ell \neq i \\ \ell \neq j}}^n \sigma_{ij, i\ell}^* S_i S_j S_\ell$$

$$(2.4.15) \quad C_{24}^* = \frac{n(n-1)(n-2)(n-3)}{K(K-1)(K-2)(K-3)} \sum_{i=1}^n \sum_{j>i}^n \sum_{\substack{\ell=1 \\ \ell \neq i \\ \ell \neq j}}^n \sum_{\substack{m>\ell \\ m \neq i \\ m \neq j}}^n \sigma_{ij, \ell m}^* S_i S_j S_\ell S_m$$

und wegen Gl. (2.2.40) können wir eine erwartungstreue Schätzfunktion $\text{var}(T_2)^*$ für $\text{var}(T_2)$ durch

$$(2.4.16) \quad \text{var}(T_2)^* = C_{22}^* + C_{23}^* + C_{24}^*$$

definieren.

Mittels Gl. (2.2.44) und (2.2.45) kann man aus C_{22}^*, C_{23}^* und C_{24}^*, σ_2^{2*} und σ_u^{2*} erwartungstreue Schätzfunktionen für die Streuungen der Erwartungswerte σ_{21}^2 und σ_{u1}^2 herleiten:

$$(2.4.17) \quad \sigma_{21}^{2*} := \sigma_2^{2*} - \frac{2(n+1)(n-2)}{n^2(n-1)^2} C_{22}^* + \frac{4}{n^2(n-1)^2} (C_{23}^* + C_{24}^*)$$

$$(2.4.18) \quad \sigma_{u1}^{2*} := \sigma_u^{2*} - \frac{2(n-2)}{n^2} C_{22}^* - \frac{n-4}{n^2} C_{23}^* + \frac{4}{n^2} C_{24}^*.$$

2.5 Schätzprobleme bei Zufallsvariablen, die sowohl einzelnen als auch Paaren von Untersuchungseinheiten zugeordnet sind

Wenn die Zufallsvariablen X_i den Untersuchungseinheiten und $\bar{X}_{ij}$ Paaren von Untersuchungseinheiten zugeordnet sind, wird bei der simultanen Schätzung von $\theta_1 = \sum_{i=1}^n X_i$ und $\theta_2 = \sum_{i=1}^n \sum_{j>i}^n \bar{X}_{ij}$ die Kovarianz der Schätzfunktionen $\hat{\theta}_1$ und $\hat{\theta}_2$ interessant. Speziell ist die Kovarianz notwendig zur Berechnung der Risikofunktion von Schätzfunktionen $\alpha\hat{\theta}_1 + \beta\hat{\theta}_2$ für einen Parameter $\theta_3 = \alpha\theta_1 + \beta\theta_2, \alpha, \beta \in \mathbb{R}$.

In diesem Abschnitt soll zunächst $\text{cov}(\hat{\theta}_1, \hat{\theta}_2)$ berechnet und erläutert werden, dann werden Schätzfunktionen der Form $\alpha\hat{\theta}_1 + \beta\hat{\theta}_2$ diskutiert und am Schluß werden erwartungstreue Schätzfunktionen für die neu eingeführten Parameter angegeben.

Die Behandlung der Probleme ist nicht so ausführlich wie bei der Schätzung von θ_1 und θ_2 , z.B. wird nicht die Zulässigkeit von Schätzfunktionen der Form $\alpha\hat{\theta}_1 + \beta\hat{\theta}_2$ untersucht und auch die Risikofunktion nicht explizit angegeben, sondern es wird nur gezeigt, wie man sie aus den bisherigen Ergebnissen und Kenntnis der Kovarianz berechnen kann. Weiterhin gelten die Voraussetzungen aus Abschnitt 2.1. Es sei allerdings angemerkt, daß $\alpha\hat{\theta}_1 + \beta\hat{\theta}_2$ keine symmetrische Schätzfunktion ist.

Wir definieren nun die zur Angabe der Kovarianz benötigten Parameter:

$$(2.5.1) \quad a_{32} := \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n E(X_i \cdot \bar{X}_{ij}) = \sum_{i=1}^n E(X_i \cdot U_i)$$

$$(2.5.2) \quad a_{33} := \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \sum_{\substack{\ell > j \\ \ell \neq i}}^n E(X_i \cdot \bar{X}_{j\ell})$$

$$(2.5.3) \quad \sigma_{3xu} := E\left(\frac{1}{n} \sum_{i=1}^n (X_i - \frac{1}{n}T_1)(U_i - \frac{2}{n}T_2)\right) \\ = \frac{1}{n}a_{32} - \frac{2}{n^2}E(T_1 \cdot T_2)$$

wobei wie bisher $U_i = \sum_{\substack{j=1 \\ j \neq i}}^n \bar{X}_{ij}$, $T_1 = \sum_{i=1}^n X_i$, $T_2 = \sum_{i=1}^n \sum_{j>i}^n \bar{X}_{ij}$ gilt.

Der erste Index 3 bei diesen Parametern soll anzeigen, daß es sich hier um Parameter zur Kovarianz der Schätzfunktionen handelt, der zweite Index bei a_{32} und a_{33} gibt die Anzahl der verschiedenen Elemente i, j, ℓ bei den Produkten $E(\bar{X}_i \cdot X_{j\ell})$ an. σ_{3xu} ist ein Maß für die Abhängigkeit der X_i und U_i , bei nicht zufälligen Werten x_i , u_i ist es die in der Stichprobentheorie gebrauchte Kovarianz bei der Variablen in der Grundgesamtheit.

Satz 2.5.1

Für die Kovarianz von $\hat{\theta}_1 = \beta(K) \sum_{i=1}^n X_i S_i$ und

$\hat{\theta}_2 = \gamma(K) \sum_{i=1}^n \sum_{j>i}^n \bar{X}_{ij} S_i S_j$ gilt:

$$\begin{aligned}
 (2.5.4) \quad \text{cov}(\hat{\theta}_1, \hat{\theta}_2) &= a_{32} E\left(\beta(K) \gamma(K) \frac{K(K-1)}{n(n-1)}\right) \\
 &+ a_{33} E\left(\beta(K) \gamma(K) \frac{K(K-1)(K-2)}{n(n-1)(n-2)}\right) \\
 &- \theta_1 \theta_2 E\left(\beta(K) \frac{K}{n}\right) E\left(\gamma(K) \frac{K(K-1)}{n(n-1)}\right) \\
 &= \sigma_{3xu} E\left(\beta(K) \gamma(K) \frac{K(K-1)}{n-1} \left[1 - \frac{K-2}{n-2}\right]\right) \\
 &+ \text{cov}(T_1, T_2) E\left(\beta(K) \gamma(K) \frac{K^2(K-1)}{n^2(n-1)}\right) \\
 &+ \theta_1 \theta_2 \text{cov}\left(\beta(K) \frac{K}{n}, \gamma(K) \frac{K(K-1)}{n(n-1)}\right),
 \end{aligned}$$

wobei a_{32} , a_{33} und σ_{3xu} in Gl. (2.5.1) - (2.5.3) definiert sind.

Beweis

Zunächst gilt

$$\begin{aligned}
 (2.5.5) \quad \text{cov}(T_1, T_2) + \theta_1 \theta_2 &= E(T_1 T_2) = E\left(\sum_{i=1}^n \sum_{j=1}^n \sum_{\ell > j}^n X_i \bar{X}_{j\ell}\right) \\
 &= a_{32} + a_{33},
 \end{aligned}$$

$$(2.5.6) \quad a_{32} = n\sigma_{3xu} + \frac{2}{n}E(T_1 T_2) \quad \text{und}$$

$$(2.5.7) \quad a_{33} = \frac{n-2}{n}E(T_1 T_2) - n\sigma_{3xu}.$$

Daraus folgt für die Kovarianz nach Gl. (1.1.28):

$$\begin{aligned}
 (2.5.8) \quad \text{cov}(\hat{\theta}_1, \hat{\theta}_2) &= E(\hat{\theta}_1 \hat{\theta}_2) - E_1(\hat{\theta}_1)E(\hat{\theta}_2) \\
 &= \sum_{i=1}^n \sum_{j=1}^n \sum_{\ell > j}^n E(X_i \bar{X}_{j\ell}) E(\beta(K) \gamma(K) S_i S_j S_\ell) \\
 &\quad - \theta_1 \theta_2 E\left(\beta(K) \frac{K}{n}\right) E\left(\gamma(K) \frac{K(K-1)}{n(n-1)}\right) \\
 &= a_{32} E\left(\beta(K) \gamma(K) \frac{K(K-1)}{n(n-1)}\right) \\
 &\quad + a_{33} E\left(\beta(K) \gamma(K) \frac{K(K-1)(K-2)}{n(n-1)(n-2)}\right) \\
 &\quad - \theta_1 \theta_2 E\left(\beta(K) \frac{K}{n}\right) E\left(\gamma(K) \frac{K(K-1)}{n(n-1)}\right) \\
 &= n\sigma_{3xu} \left[E\left(\beta(K) \gamma(K) \frac{K(K-1)}{n(n-1)}\right) - E\left(\beta(K) \gamma(K) \frac{K(K-1)(K-2)}{n(n-1)(n-2)}\right) \right] \\
 &\quad + \text{cov}(T_1, T_2) \left[\frac{2}{n} E\left(\beta(K) \gamma(K) \frac{K(K-1)}{n(n-1)}\right) \right. \\
 &\quad \left. + \frac{n-2}{n} E\left(\beta(K) \gamma(K) \frac{K(K-1)(K-2)}{n(n-1)(n-2)}\right) \right] \\
 &\quad + \theta_1 \theta_2 \left[\frac{2}{n} E\left(\beta(K) \gamma(K) \frac{K(K-1)}{n(n-1)}\right) \right. \\
 &\quad \left. + \frac{n-2}{n} E\left(\beta(K) \gamma(K) \frac{K(K-1)(K-2)}{n(n-1)(n-2)}\right) \right. \\
 &\quad \left. - E\left(\beta(K) \frac{K}{n}\right) E\left(\gamma(K) \frac{K(K-1)}{n(n-1)}\right) \right]
 \end{aligned}$$

Ein Ausrechnen der Faktoren ergibt den 2. Teil von Gl.(2.5.4).

Bemerkung 2.5.2

Für die aus dem arithmetischen Mittel der Stichprobenwerte

berechneten Schätzfunktionen θ_1^* und θ_2^* erhalten wir:

$$\begin{aligned} (2.5.9) \quad \text{cov}(\theta_1^*, \theta_2^*) &= na_{32}E\left(\frac{1}{K}\right) + \frac{n}{n-2}a_{33}\left(1-2E\left(\frac{1}{K}\right)\right) - \theta_1\theta_2 \\ &= \frac{n^3}{n-2}\left(E\left(\frac{1}{K}\right) - \frac{1}{n}\right)\sigma_{3xu} + \text{cov}(T_1, T_2) \end{aligned}$$

Für die Horvitz-Thompson-Schätzfunktionen $\tilde{\theta}_1$ und $\tilde{\theta}_2$ gilt:

$$\begin{aligned} (2.5.10) \quad \text{cov}(\tilde{\theta}_1, \tilde{\theta}_2) &= a_{32} \frac{1}{p_1} + a_{33} \frac{p_3}{p_1 p_2} - \theta_1 \theta_2 \\ &= \sigma_{3xu} n \left(\frac{1}{p_1} - \frac{p_3}{p_1 p_2} \right) \\ &\quad + \text{cov}(T_1, T_2) \left(\frac{(n-2)p_3}{np_1 p_2} + \frac{2}{np_1} \right) \\ &\quad + \theta_1 \theta_2 \left(\frac{(n-2)p_3}{np_1 p_2} + \frac{2}{np_1} - 1 \right) . \end{aligned}$$

Bemerkung 2.5.3

Bei den Streuungen σ_1^2 und σ_2^2 haben wir jeweils zusätzlich die Streuungen σ_{11}^2 und σ_{21}^2 der Mittelwerte definiert und mit σ_1^2 , σ_2^2 in Verbindung gesetzt. Wir wollen nun das gleiche für die Kovarianz durchführen. Die Kovarianz der Mittelwerte $E(X_i)$ und $E(U_i)$ in der Grundgesamtheit lautet:

$$\begin{aligned}
 (2.5.11) \quad \sigma_{3xu1} &:= \frac{1}{n} \sum_{i=1}^n (E(X_i) - \frac{1}{n}\theta_1) (E(U_i) - \frac{2}{n}\theta_2) \\
 &= \frac{1}{n} \sum_{i=1}^n E(X_i)E(U_i) - \frac{2}{n^2}\theta_1\theta_2 \\
 &= \frac{1}{n}a_{32} - \frac{1}{n} \sum_{i=1}^n \text{cov}(X_i, U_i) - \frac{2}{n^2}[E(T_1T_2) - \text{cov}(T_1T_2)] \\
 &= \sigma_{3xu} + \frac{2}{n^2} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \sum_{\substack{\ell > j \\ \ell \neq i}}^n \text{cov}(X_i, \bar{X}_{j\ell}) \\
 &\quad - \frac{n-2}{n^2} \sum_{i=1}^n \text{cov}(X_i, U_i)
 \end{aligned}$$

Wenn wir $\text{cov}(\hat{\theta}_1, \hat{\theta}_2)$ mittels σ_{3xu1} ausdrücken wollen, müssen wir also in Gl. (2.5.4) σ_{3xu} wie folgt ersetzen:

$$\begin{aligned}
 (2.5.12) \quad \sigma_{3xu} &= \sigma_{3xu1} + \frac{n-2}{n^2} \sum_{i=1}^n \text{cov}(X_i, U_i) \\
 &\quad - \frac{2}{n^2} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \sum_{\substack{\ell > j \\ \ell \neq i}}^n \text{cov}(X_i, \bar{X}_{j\ell})
 \end{aligned}$$

Wegen:

$$\begin{aligned}
 (2.5.13) \quad \text{cov}(T_1, T_2) &= \sum_{i=1}^n \text{cov}(X_i, U_i) \\
 &\quad + \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \sum_{\substack{\ell > j \\ \ell \neq i}}^n \text{cov}(X_i, \bar{X}_{j\ell}) ,
 \end{aligned}$$

kann man in (2.5.4) dann noch einen der Parameter
 $\text{cov}(T_1, T_2), \quad \sum_{i=1}^n \text{cov}(X_i, U_i)$ oder $\sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \sum_{\substack{\ell > j \\ \ell \neq i}}^n \text{cov}(X_i, \bar{X}_{j\ell})$
 durch die beiden anderen darstellen.

In den folgenden zwei Bemerkungen wollen wir auf die Schätzungen von Linearkombinationen $\theta_3 = \alpha\theta_1 + \beta\theta_2$; $\alpha, \beta \in \mathbb{R}$ von θ_1 und θ_2 eingehen. Probleme dieser Art kann man bei der Inferenz aus Graphen (Kapitel III) finden; ein anderes Beispiel hatten wir in Abschnitt 1.5 bei der Schätzung von $\sigma_1^2 = \frac{n-1}{n^2} a_{11} - \frac{1}{n^2} a_{12}$, wo a_{11} und a_{12} Erwartungswerte von Summen von Zufallsvariablen x_i^2 und $\bar{x}_{ij} = x_i x_j$ waren, die den Untersuchungseinheiten bzw. Paaren von Untersuchungseinheiten zugeordnet waren. Mit den folgenden Ergebnissen kann man nun die Varianz der Schätzfunktion σ_1^{2*} bestimmen, was allerdings hier nicht durchgeführt werden soll.

Bemerkung 2.5.4

Es liegt nahe, $\theta_3 = \alpha\theta_1 + \beta\theta_2$ durch die entsprechende Linearkombination der Schätzfunktionen $\hat{\theta}_3 = \alpha\hat{\theta}_1 + \beta\hat{\theta}_2$ zu schätzen. Falls die Koeffizienten $\beta(K)$ bei $\hat{\theta}_1$ und $\gamma(K)$ bei $\hat{\theta}_2$ verschieden sind, ist allerdings die Schätzfunktion $\hat{\theta}_3 = \alpha\hat{\theta}_1 + \beta\hat{\theta}_2$ nicht symmetrisch, weil beobachtete Werte x_i und $\bar{x}_{ij}$ verschieden gewichtet werden. Insbesondere sind die Schätzfunktionen $\theta_3^* = \alpha\theta_1^* + \beta\theta_2^*$, $\tilde{\theta}_3 = \alpha\tilde{\theta}_1 + \beta\tilde{\theta}_2$ oder z.B. σ_1^{2*} nicht symmetrisch. Symmetrische lineare Schätzfunktionen für θ_3 haben die Form:

$$(2.5.14) \quad \hat{\theta}_3 = \delta(K) \sum_{i=1}^n \sum_{j \geq i}^n \bar{x}_{ij} s_i s_j$$

mit $\bar{X}_{ii} = X_i$ und es gilt:

$$(2.5.15) \quad E(\hat{\theta}_3) = \alpha\theta_1 E\left(\delta(K)\frac{K}{n}\right) + \beta\theta_2 E\left(\delta(K)\frac{K(K-1)}{n(n-1)}\right)$$

Der Erwartungswert hängt nicht nur von θ_3 , sondern auch von dem Verhältnis von θ_1 und θ_2 ab, z.B. gilt für $\theta_2 = 0$:

$$(2.5.16) \quad E(\hat{\theta}_3) = \alpha\theta_1 E\left(\delta(K)\frac{K}{n}\right) = \theta_3 E\left(\delta(K)\frac{K}{n}\right)$$

und für $\theta_1 = 0$:

$$(2.5.17) \quad E(\hat{\theta}_3) = \beta\theta_2 E\left(\delta(K)\frac{K(K-1)}{n(n-1)}\right) = \theta_3 E\left(\delta(K)\frac{K(K-1)}{n(n-1)}\right)$$

Falls also das Verhältnis von θ_1 und θ_2 nicht bekannt ist, kann man kein $\delta(K)$ finden, daß in Gl. (2.5.16) und (2.5.17) gleichzeitig $E(\hat{\theta}_3) = \theta_3$ gilt und somit enthält die Klasse der linearen symmetrischen Schätzfunktionen der Form

(2.5.14) im allgemeinen keine erwartungstreuen Schätzfunktionen. Es ist also nicht sinnvoll, sich auf diese Klasse zu beschränken. Einen wichtigen Fall, in dem das Verhältnis von θ_1 und θ_2 bekannt ist, werden wir in Abschnitt 3.6 behandeln, wenn wir annehmen, daß für alle Paare von Untersuchungseinheiten der Stichprobe die Realisierungen $\bar{X}_{ij}$ beobachtet werden und zusätzlich zu jeder Untersuchungseinheit der Stichprobe der Wert von

$$U_i = \sum_{\substack{j=1 \\ j \neq i}}^n \bar{X}_{ij} \text{ beobachtet wird. Für } \theta_1 = E\left(\sum_{i=1}^n U_i\right)$$

gilt in diesem Fall $\theta_1 = 2\theta_2 = 2E\left(\sum_{i=1}^n \sum_{j>i}^n \bar{X}_{ij}\right)$ und man kann

erwartungstreue Schätzfunktionen der Form (2.5.14) konstruieren.

Bemerkung 2.5.5

In Satz 1.3.3 bzw. 1.3.5 haben wir $E(\hat{\theta}_1 - \theta_1)^2$, in Satz 2.2.1 bzw. 2.2.5 $E(\hat{\theta}_2 - \theta_2)^2$ und in Satz 2.5.1 $\text{cov}(\hat{\theta}_1, \hat{\theta}_2)$ bestimmt. Mit Hilfe dieser Ergebnisse können wir $E(\hat{\theta}_3 - \theta_3)^2$ mit $\theta_3 = \alpha\theta_1 + \beta\theta_2$ und $\hat{\theta}_3 = \alpha\hat{\theta}_1 + \beta\hat{\theta}_2$ nach folgender Gleichung berechnen:

$$\begin{aligned} (2.5.18) \quad E(\hat{\theta}_3 - \theta_3)^2 &= \alpha^2 E(\hat{\theta}_1 - \theta_1)^2 + \beta^2 E(\hat{\theta}_2 - \theta_2)^2 \\ &\quad + 2\alpha\beta E((\hat{\theta}_1 - \theta_1)(\hat{\theta}_2 - \theta_2)) \\ &= \alpha^2 E(\hat{\theta}_1 - \theta_1)^2 + \beta^2 E(\hat{\theta}_2 - \theta_2)^2 \\ &\quad + 2\alpha\beta \text{cov}(\hat{\theta}_1, \hat{\theta}_2) + 2\alpha\beta (E(\hat{\theta}_1) - \theta_1)(E(\hat{\theta}_2) - \theta_2) \end{aligned}$$

Da dieser Erwartungswert von 9 Parametern abhängt (σ_1^2 , σ_2^2 , σ_u^2 , σ_{3xu} , $\text{var}(T_1)$, $\text{var}(T_2)$, $\text{cov}(T_1, T_2)$, θ_1 , θ_2) geben wir die Formel nicht explizit an, sondern wollen nur $E((\hat{\theta}_1 - \theta_1)(\hat{\theta}_2 - \theta_2))$ als Funktion dieser Parameter darstellen, wobei sich die $\beta(K)$ wie bisher auf $\hat{\theta}_1$ und $\gamma(K)$ auf $\hat{\theta}_2$ beziehen:

$$\begin{aligned} (2.5.19) \quad E((\hat{\theta}_1 - \theta_1)(\hat{\theta}_2 - \theta_2)) &= \sigma_{3xu} E\left(\beta(K)\gamma(K)\frac{K(K-1)}{n-1}\right) \left(1 - \frac{K-2}{n-2}\right) \\ &\quad + \text{cov}(T_1, T_2) E\left(\beta(K)\gamma(K)\frac{K^2(K-1)}{n^2(n-1)}\right) \\ &\quad + \theta_1\theta_2 E\left[\left(\beta(K)\frac{K}{n} - 1\right)\left(\gamma(K)\frac{K(K-1)}{n(n-1)} - 1\right)\right] \end{aligned}$$

Bemerkung 2.5.6

Wir wollen erwartungstreue Schätzfunktionen für a_{32} , a_{33} und σ_{3xu} angeben:

$$(2.5.20) \quad a_{32}^* := \frac{n(n-1)}{K(K-1)} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n x_i \bar{x}_{ij} s_i s_j$$

$$(2.5.21) \quad a_{33}^* := \frac{n(n-1)(n-2)}{K(K-1)(K-2)} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \sum_{\substack{\ell > j \\ \ell \neq i}}^n x_i \bar{x}_{j\ell} s_i s_j s_\ell$$

$$(2.5.22) \quad \sigma_{3xu}^* := \frac{n-2}{n} a_{32}^* - \frac{1}{n} a_{33}^*$$

Aus der Unabhängigkeit der Werte vom Stichprobenverfahren und aus Gl. (1.1.28) folgt die Erwartungstreue der Schätzfunktion a_{32}^* und a_{33}^* ; aus Gl. (2.5.3) und (2.5.5) folgt die von σ_{3xu}^* .

Bemerkung 2.5.7

Um die Kovarianz von T_1 und T_2 zu schätzen, sind aus den in Abschnitt 1.5 und 2.4 besprochenen Gründen mehrere Beobachtungen zu jeder Untersuchungseinheit der Stichprobe notwendig. X_i und $\bar{x}_{ij}$ sei wie in Gl. (1.5.10) und (2.4.10) als Mittel der κ unabhängigen, identisch verteilten Variablen $X_{iv}, \bar{x}_{ijv}, v=1, \dots, \kappa$ definiert. Die Kovarianz $\sigma_{i,j\ell}$ von X_i und $\bar{x}_{j\ell}$ ($j \neq \ell$) kann man nach Stenger (1971, S. 33) erwartungstreu durch:

$$(2.5.23) \quad \sigma_{i,j\ell}^* := \frac{1}{K(K-1)} \sum_{v=1}^K (X_{iv} - \bar{X}_i) (\bar{X}_{ijv} - \bar{X}_{ij})$$

schätzen und nach Gl. (2.5.13) und (2.5.5) erhält man als erwartungstreue Schätzfunktionen für $\text{cov}(T_1, T_2)$ und $\theta_1 \cdot \theta_2$:

$$(2.5.24) \quad \begin{aligned} \text{cov}(T_1, T_2)^* &= \frac{n(n-1)}{K(K-1)} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \sigma_{i,ij}^* S_i S_j \\ &+ \frac{n(n-1)(n-2)}{K(K-1)(K-2)} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \sum_{\substack{\ell > j \\ \ell \neq i}}^n \sigma_{i,j\ell}^* S_i S_j S_\ell \end{aligned}$$

$$(2.5.25) \quad (\theta_1 \theta_2)^* = a_{32}^* + a_{33}^* - \text{cov}(T_1, T_2)^*$$

Mit Hilfe der in den beiden letzten Bemerkungen gebrachten Schätzfunktionen kann man nun $\text{cov}(\hat{\theta}_1, \hat{\theta}_2)$ schätzen und mit den Ergebnissen aus Abschnitt 1.5 und 2.4 kann man zusätzlich $E(\hat{\theta}_3 - \theta_3)^2$ für die in Bemerkung 2.5.4 und 2.5.5 besprochenen Linearkombinationen von θ_1 und θ_2 schätzen.

III. ANWENDUNG DER ERGEBNISSE AUF INFERENZ AUS GRAPHEN
MIT BESONDERER BERÜCKSICHTIGUNG SOZIALWISSEN-
SCHAFTLICHER FRAGESTELLUNGEN

3.1 Graphen als Modell in den Sozialwissenschaften

3.1.1 Einführung

Eigenschaften, die sich auf Paare bzw. geordnete Tupel von Einheiten beziehen, kann man mit dem mathematischen Modell des Graphen oder Netzwerks darstellen und behandeln. Gerade in den Sozialwissenschaften haben Graphen und Netzwerke eine große Rolle bei der Darstellung von sozialen Beziehungen bekommen; hier bezieht sich der Begriff des sozialen Netzwerks "auf das Geflecht der - in der Regel: sozialen - Beziehungen, die zwischen einer definierten Menge von einzelnen Einheiten - in der Regel: Individuen - beobachtet werden können" (Kähler (1975)). In diesem Abschnitt werden einige grundlegende Begriffe der Graphentheorie eingeführt, an Beispielen erklärt und am Schluß wird ein Einblick in die umfangreiche Literatur zur Analyse sozialer Netzwerke gegeben. Zufallsgraphen, Inferenz aus Graphen und die zugehörige Literatur werden allerdings erst ab Abschnitt 3.2 behandelt. Wenn wir uns auch auf Anwendungen der Graphentheorie in den Sozialwissenschaften beschränken wollen, muß ihre Bedeutung in der Informatik, Ingenieurwissenschaften und Operations Research genannt werden; für mathematisch interessierte Leser sei auf die bekannten Monographien von Berge (1962) und Harary (1969) verwiesen, in denen allerdings andere Schwerpunkte als in dieser Arbeit und bei der Analyse sozialer Netzwerke gesetzt werden.

Die folgende Darstellung der Grundbegriffe richtet sich im wesentlichen nach Frank (1981), (1971), die deutschen Begriffe sind den Büchern von Kliemann/Müller (1976) und Biess (1976) entnommen.

3.1.2 Ungerichtete Graphen

Ungerichtete Graphen sind ein Modell zur Darstellung wechselseitiger Beziehungen zwischen Objekten. Formal besteht ein ungerichteter (einfacher) Graph $G = (\Omega, M)$ aus einer nicht leeren Menge Ω von Knoten (oder Punkten; englisch: nodes, vertices oder points) und einer Menge $M \subseteq \{\{i, j\} : i \in \Omega, j \in \Omega, i \neq j\}$ von verschiedenen, ungeordneten Paaren von Knoten, den Kanten (Linien; englisch: edges, links, lines). Zu einer Kante $\{i, j\}$ heißen i und j Endpunkte. Knoten i und j heißen adjacent, wenn sie durch eine Kante verbunden werden, d.h. $\{i, j\} \in M$. Wir werden uns ausschließlich mit Graphen mit endlicher Knotenmenge $\Omega = \{1, 2, \dots, n\}$ beschäftigen, n ist dann die Ordnung (order) von G , die Anzahl der Kanten $|M|$ heißt auch Größe (size) von G . Die Größe eines ungerichteten Graphen kann höchstens $\binom{n}{2}$ betragen; in diesem Fall sind alle Knoten durch Kanten verbunden und der Graph heißt vollständig (complete).

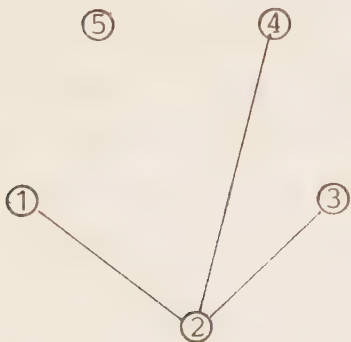


Abb. 3.1.1

Anschaulich werden ungerichtete Graphen dargestellt, indem man Knoten als Punkte oder Kreise und Kanten als Verbindungslinien zeichnet. Die nebenstehende Abb. 3.1.1 zeigt einen Graphen mit Knotenmenge $\Omega = \{1, 2, 3, 4, 5\}$ und Kantenmenge $M = \{\{1, 2\}, \{2, 3\}, \{2, 4\}\}$. Bei sozialen Netzwerken repräsentieren die Knoten oft Personen, die Kanten

symmetrische Beziehungen zwischen Personen (z.B. Personen i und j kennen sich gegenseitig, haben sich in einem bestimmten

Zeitraum getroffen oder besucht usw.). Kommunikationsstrukturen innerhalb von Organisationen oder aufgabenorientierten Gruppen werden durch Graphen dargestellt, bei denen Knoten Personen oder Positionen (z.B. Direktor, Abteilungsleiter, Sachbearbeiter usw.) bedeuten und die Kanten durch die Organisationsform vorgegebene Kommunikationswege zwischen den Einheiten aufzeigen. Andere Anwendungsmöglichkeiten liegen vor, wenn man Verkehrsverbindungen zwischen Städten, politische Beziehungen zwischen Staaten (Handels-, militärische Verträge) oder Beziehungen zwischen Personengruppen betrachtet.

3.1.3 Gerichtete Graphen

Bei der Untersuchung nicht notwendig wechselseitiger Beziehungen zwischen Objekten sind gerichtete Graphen $G = (\Omega, M)$ (Digraphen) geeignete Modelle, die aus einer nicht leeren Menge Ω von Knoten und einer Menge $M \subseteq \{(i, j) : i \in \Omega, j \in \Omega, i \neq j\}$ von geordneten Tupeln von verschiedenen Knoten bestehen: den (gerichteten) Kanten (Bögen; englisch: arcs, directed edges). Zu einer gerichteten Kante (i, j) heißt i Anfangs- und j Endknoten, i heißt adjazent nach (adjacent to) j bzw. j heißt adjazent von (adjacent from) i , wenn $(i, j) \in M$ gilt. Die maximale Größe (d.h. Anzahl der Kanten) bei einem gerichteten Graphen mit n Knoten beträgt $n(n-1)$, in diesem Fall heißt der Graph vollständig und alle Knotenpaare sind in beide Richtungen durch Kanten

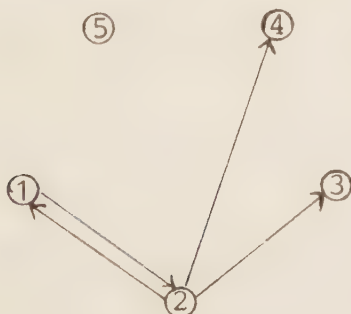


Abb. 3.1.2

verbunden. Anschaulich werden bei gerichteten Graphen die Kanten durch Pfeile dargestellt, die nebenstehende Abb. 3.1.2 zeigt einen gerichteten Graphen mit Knotenmenge $\Omega = \{1, 2, 3, 4, 5\}$ und Kantenmenge $\{(1, 2), (2, 1), (2, 3), (2, 4)\}$.

Gerichtete Graphen erlauben eine feinere Darstellung von Strukturen, da sie wechselseitige Beziehungen zwischen Einheiten durch Kanten in beide Richtungen darstellen können und bei nicht wechselseitiger Beziehung durch die Richtung der Kanten die Richtung der Beziehung angeben können. Gerade bei empirischen sozialen Netzwerken liegen meistens nichtsymmetrische Beziehungen vor; das klassische Beispiel der Soziometrie, die Wahl oder Ablehnung von Personen in einer definierten Gruppe, liefert nicht notwendig wechselseitige Beziehungen, und gerade die Unsymmetrie dieser Beziehungen ist wichtiger Untersuchungsgegenstand der Soziometrie. Wenn z.B. in einer Schulklasse ein Schüler i den Schüler j als Sitznachbarn wählen würde, impliziert das nicht, daß j neben i sitzen möchte. Bei einer Untersuchung der Klassenstruktur ist meistens die Unterscheidung von wechselseitigen und einseitigen Wahlen wichtig. Die Berücksichtigung der Richtung kann auch bei Kommunikationsstrukturen nützlich sein; z.B. kann möglicherweise ein Leiter einer Organisation mit allen Untergebenen kommunizieren, während Untergebene einen bestimmten Instanzenweg einhalten müssen. Bei der empirischen Beobachtung von Kommunikationsvorgängen oder Kontakten (Besuche, Telefongespräche usw.) wird normalerweise die Richtung dieses Kontaktes mit beobachtet (wer besucht und wer wird besucht, von wem geht die Kommunikation aus). Außerdem eignen sich gerichtete Graphen zur Modellierung von Machtstrukturen, wobei eine Kante (i,j) bedeutet, daß die Person i in einem spezifizierten Sinne Macht (z.B. Weisungsbefugnis in einer Organisation) über Person j hat.

3.1.4 Darstellung von Graphen mittels Kantenindikatormatrizen

Für Berechnungen oder z.B. zur Eingabe in einen Computer ist es sinnvoll, endliche Graphen der Ordnung n

durch $n \times n$ Matrizen darzustellen. Zu einem gerichteten oder ungerichteten Graphen heißt die Matrix y mit den Elementen

$$y_{ij} := \begin{cases} 1 & \text{falls } (i,j) \in M \text{ bzw. } \{i,j\} \in M \\ 0 & \text{sonst} \end{cases}$$

Kantenindikator- oder Adjazenzmatrix. Die Elemente $y_{ii}, i \in \Omega$ in der Diagonalen betragen 0, wegen $\{i,j\} = \{j,i\}$ ist die Matrix bei ungerichteten Graphen symmetrisch. Zu den in Abb. 3.1.1 und 3.1.2 gezeichneten Graphen lauten die Kantenindikatormatrizen:

$$y_1 = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad y_2 = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Im Gegensatz zur üblichen Bezeichnung von Matrizen mit Großbuchstaben haben wir kleine Buchstaben gewählt, um die im nächsten Abschnitt eingeführten Zufallsmatrizen von ihren Realisierungen zu unterscheiden.

3.1.5 Subgraphen, Teilgraphen, Isomorphie, Zusammenhang

Zu einem Graph $G = (\Omega, M)$ besteht ein Subgraph G' aus einer Teilmenge Ω' der Knoten und allen Kanten die Knoten dieser Teilmenge verbinden, d.h. $G' = (\Omega', \Omega' \times \Omega' \cap M)$. Ein Teilgraph G' (partial Graph) besteht aus allen Knoten des ursprünglichen Graphen und einer Teilmenge $M' \subset M$ der Kanten, d.h. $G = (\Omega, M')$. Wir werden später Sub- oder Teilgraphen als Stichprobengraphen behandeln (Abschnitt 3.3).

Bei einigen soziometrischen Untersuchungen interessieren Eigenschaften bestimmter Knoten; wenn man in einer Schulklasse die Schüler nach dem gewünschten Sitznachbarn fragt, will man erfahren, welcher Schüler (durch einen bestimmten Knoten repräsentiert) am beliebtesten ist, also am meisten

gewählt wird oder welche Schüler überhaupt nicht gewählt werden. Falls man dagegen die Struktur einer Gruppe untersuchen will, also z.B. in der Schulklasse nicht wissen will, wer am beliebtesten ist, sondern nur wissen will, ob jemand ganz besonders häufig gewählt wird und was für ein Anteil überhaupt nicht gewählt wird, betrachtet man Eigenschaften, die unabhängig von der Bezeichnung bzw. Numerierung der Knoten sind. Graphen gleicher Ordnung, die sich nur durch die Bezeichnung der Knoten unterscheiden, heißen isomorph. Gerade bei statistischen oder empirischen Problemen ist man hauptsächlich an Eigenschaften interessiert, die nicht von der Bezeichnung der Knoten abhängen. Wir werden deshalb isomorphe Graphen als identisch ansehen und Mengen isomorpher Graphen durch ein Element repräsentieren. Für die Kantenindikatormatrix bedeutet Isomorphie, daß nur die Reihenfolge der Zeilen und in gleicher Weise die Reihenfolge der Spalten geändert werden kann, z.B. sind zu den drei folgenden Matrizen die Graphen isomorph.

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

vgl. Abb. 3.1.2

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

Vertauschung der
Nummern von
Knoten 1 und 5

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

zusätzlich Ver-
tauschung von
Knoten 2 und 3

Bei der Analyse sozialer Netzwerke werden oft Untergruppen gesucht, die eng bzw. überhaupt durch Kontakte (Beziehungen) verbunden sind. Um einen kleinen Einblick zu geben, seien hier die Konzepte der "Clique" und "Zusammenhangskomponente" genannt. Eine Clique ist ein maximaler, vollständiger Subgraph; alle geordneten bzw. ungeordneten Knotenpaare einer Clique sind also Kanten. Zwei Knoten i und j heißen in ungeordneten Graphen zusammenhängend, wenn es eine Kantenfolge $\{i, i_1\}, \{i_1, i_2\}, \dots, \{i_{k-1}, j\}$ gibt, die i und j verbindet;

z.B. in Kommunikationsnetzwerken ist also eine Kommunikation von i nach j direkt oder über Vermittler möglich. Bei gerichteten Graphen muß man verschiedene Arten des Zusammenhangs je nach den Richtungen der Kanten der Verbindung unterscheiden (schwach, stark, unilateral zusammenhängend). Die zusammenhängenden Knoten bilden Subgraphen, die (Zusammenhangs-) Komponenten eines Graphen.

3.1.6 Parameter von ungerichteten Graphen

Wir gehen von Graphen mit fester Knotenmenge $\Omega = \{1, 2, \dots, n\}$ aus und bezeichnen bei ungerichteten Graphen die Elemente der Kantenindikatormatrix mit $\bar{y}_{ij}$, um sie von den Matrixelementen y_{ij} bei gerichteten Graphen zu unterscheiden (vgl. Notation in Abschnitt 2.1).

Ein Parameter zur Beschreibung von Graphen wurde schon genannt, die Größe

$$(3.1.1) \quad t_{\bar{y}2} := |M| = \sum_{i=1}^n \sum_{j>i}^n \bar{y}_{ij} .$$

Der Grad g_i eines Knotens i ist die Anzahl der Kanten, die einen Endpunkt in i haben.

$$(3.1.2) \quad g_i := \sum_{\substack{j=1 \\ j \neq i}}^n \bar{y}_{ij} ;$$

er ist die i -te Zeilen- bzw. Spaltensumme der (symmetrischen) Kantenindikatormatrix.

Die Häufigkeitsverteilung der Knotengrade bzw. Parameter dieser Häufigkeitsverteilung geben Auskunft über die Art des zugrundeliegenden Graphen. Für den Mittelwert der Grade gilt:

$$(3.1.3) \quad \frac{1}{n} \sum_{i=1}^n g_i := \frac{2}{n} t_{\bar{y}2} \quad ;$$

$t_{\bar{y}2}$ ist also ein Parameter der Häufigkeitsverteilung der Grade; ein anderer wichtiger Parameter dieser Verteilung ist die Varianz s_g^2 der Knotengrade als Maß der Heterogenität der Knoten.

$$(3.1.4) \quad s_g^2 := \frac{1}{n} \sum_{i=1}^n (g_i - \frac{2}{n} t_{\bar{y}2})^2$$

Bei sozialen Netzwerken gibt $t_{\bar{y}2}$ Kenntnis über die Häufigkeit der durch Kanten des Graphen dargestellten sozialen Beziehung; abgeleitet aus $t_{\bar{y}2}$ ist der Index der sozialen Dichte (Density) $\frac{2}{n(n-1)} t_{\bar{y}2}$ als der Anteil der beobachteten Beziehungen zwischen Individuen oder anderen Untersuchungseinheiten an der Anzahl aller theoretisch möglichen Beziehungen. Der Mittelwert der Grade $\frac{2}{n} t_{\bar{y}2}$ ist in der Anwendung häufig die durchschnittliche Größe des Bekannten- oder Freundeskreises (average acquaintance volume) von Personen einer Population (vgl. Granovetter (1976)). Die Varianz der Knotengrade mißt dann die Streuung der durchschnittlichen Größe des Bekanntenkreises oder, wenn die Kanten Kontakte zwischen Individuen bedeuten, die Streuung der Zahl der Kontakte der Individuen.

Häufig werden Graphen durch die Anzahl bestimmter Typen von Subgraphen, speziell Subgraphen der Ordnung 2 und 3, beschrieben. Bis auf Isomorphie gibt es 2 Graphen der Ordnung 2 und 4 Graphen der Ordnung 3; sie sind in der folgenden Abbildung dargestellt:

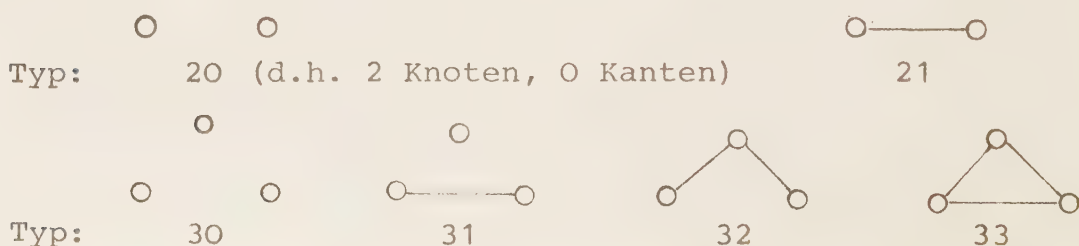


Abb. 3.1.3

Subgraphen der Ordnung 2 heißen Dyaden, der Ordnung 3 Triaden. Zu einem Graphen der Ordnung n gibt es $\binom{n}{2}$ Dyaden und $\binom{n}{3}$ Triaden und die "dyad-" bzw. "triad counts" $t_{\ell m}$ geben die Anzahl der obengenannten Typen von Dyaden bzw. Triaden an: t_{20} ist die Anzahl der Dyaden vom Typ 20, t_{21} die Anzahl vom Typ 21 usw. Mit den dyad- und triad counts lassen sich Graphen sehr gut beschreiben, aus ihnen kann man z.B. $t_{\bar{y}2}$ und die Varianz der Knotengrade berechnen oder die Anzahl der Zusammenhangskomponenten nach unten abschätzen (vgl. Frank (1981)); neben der Dichte kann man z.B. den Anteil der Triaden vom Typ 33 an allen $\binom{n}{3}$ Triaden als Maß des Zusammenhangs, der "strukturellen Kohäsion", einer Gruppe benutzen (Friedkin (1981)). Obwohl im Rahmen dieser Arbeit die dyad- und triad counts keine wesentliche Rolle spielen, sind sie hier wegen ihrer großen Bedeutung in der Literatur eingeführt und es wird gelegentlich auf sie verwiesen.

3.1.7 Parameter von gerichteten Graphen

Es sei $\Omega = \{1, 2, \dots, n\}$ und $y_{ij}, i \in \Omega, j \in \Omega$ seien die Elemente der Kantenindikatormatrix. Die Größe t_{y2} des gerichteten Graphen beträgt

$$(3.1.5) \quad t_{y2} := |M| = \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n y_{ij}.$$

Bei den Graden von Knoten müssen wir nun zwischen Ausgangsgrad und Eingangsgrad unterscheiden; der Ausgangsgrad a_i eines Knotens i ist die Anzahl der Kanten mit Anfangsknoten i

$$(3.1.6) \quad a_i := \sum_{\substack{j=1 \\ j \neq i}}^n y_{ij},$$

der Eingangsgrad e_i ist die Anzahl der Kanten mit Endpunkt im Knoten i

$$(3.1.7) \quad e_i := \sum_{\substack{j=1 \\ j \neq i}}^n y_{ji} .$$

Bezogen auf die Kantenindikatormatrix sind die Ausgangsgrade Zeilensummen und die Eingangsgrade Spaltensummen. Es gilt:

$$(3.1.8) \quad \sum_{i=1}^n a_i = \sum_{i=1}^n e_i = t_{y2} ;$$

die Summe und damit auch die Mittel der Ausgangs- und Eingangsgrade sind also gleich. Weitere Parameter der Häufigkeitsverteilung der Aus- oder Eingangsgrade sind die Varianzen s_a^2 und s_e^2 , die im Gegensatz zu den Mitteln der Aus- und Eingangsgrade verschieden sein können:

$$(3.1.9) \quad s_a^2 := \frac{1}{n} \sum_{i=1}^n (a_i - \frac{1}{n} t_{y2})^2$$

$$(3.1.10) \quad s_e^2 := \frac{1}{n} \sum_{i=1}^n (e_i - \frac{1}{n} t_{y2})^2 .$$

Die Dichte als Anteil der vorhandenen Kanten an allen theoretisch möglichen Kanten beträgt bei gerichteten Graphen $\frac{1}{n(n-1)} t_{y2}$.

Wir wollen wieder auf die Beschreibung von Graphen mittels "dyad-" und "triad counts" hinweisen; da die Richtung der Kanten eine Rolle spielt, gibt es mehr verschiedene Typen von Subgraphen der Ordnung 2 und 3. Die drei verschiedenen Arten von Dyaden sind in der folgenden Abbildung dargestellt, und außerdem gibt es 16 Arten von Triaden.

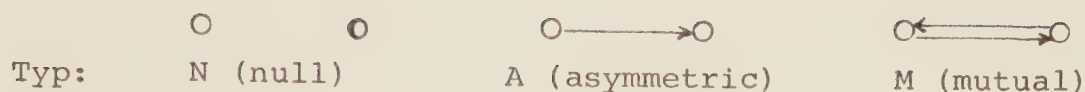


Abb. 3.1.4

Die Dyaden vom Typ M spielen eine wichtige Rolle, ihre Anzahl ist die Summe der wechselseitigen Beziehungen (mutual arcs) zwischen den Untersuchungseinheiten. Wenn man nur die wechselseitigen Beziehungen untersucht, ist es sinnvoll, den aus diesen Beziehungen gebildeten ungerichteten Graphen zu betrachten.

Bei empirisch beobachteten sozialen Netzwerken, insbesondere in der Soziometrie, liegen in der Regel gerichtete Graphen vor; die Bedeutung der Parameter soll deshalb im folgenden Abschnitt an soziometrischen Fragestellungen erläutert werden.

3.1.8 Einige Indizes der Soziometrie

Der wichtigste Vertreter der Soziometrie, Moreno, stellte 1953 die Soziometrie als eigenständige Wissenschaft der Soziologie und dem wissenschaftlichen Sozialismus gegenüber. Später bezeichnet er sie allerdings nur noch als Wissenschaft zur Messung zwischenmenschlicher Beziehungen (vgl. Moreno (1974) S. XXV, 385). Die Soziometrie wird heute noch etwas einschränken-der als eine Methode angesehen, "um spezielle Erscheinungsformen der Interaktion von Individuen zu untersuchen" (Nehnevajsa (1973) S. 260). In einem sozialen System werden dabei hauptsächlich die Strukturen untersucht, die auf informellen Beziehungen beruhen; es werden z.B. Kontakte von Personen beobachtet oder versucht, Zu- und Abneigung der Personen festzustellen. Nach Nehnevajsa, auf den wir uns im folgenden beziehen, werden soziometrische Forschungen heute seltener berichtet als in den fünfziger Jahren, soziometrische oder

soziometrieähnliche Daten tauchen allerdings in einer Vielzahl von Studien auf, wobei der allgemeinere Begriff Netzwerkanalyse die Bezeichnung Soziometrie in den Hintergrund gedrängt hat.

Soziometrische Daten erhält man, indem man durch Beobachtung oder Befragung der Mitglieder einer Gruppe positive, negative oder indifferente (bzw. neutrale) Beziehungen ermittelt. Ein häufiges Beispiel hierfür ist die Befragung der Schüler einer Klasse nach den Mitschülern, neben denen sie sitzen möchten und den Schülern, neben denen sie nicht sitzen möchten. Wenn die Wahl bzw. Ablehnung des Schülers j durch Schüler i als gerichtete Kante von i nach j aufgefaßt wird, erhält man als Modell einen gerichteten Graphen. Die zeichnerische Darstellung (siehe Abb. 3.1.1 und 3.1.2) des Graphen heißt im Rahmen der Soziometrie Soziogramm, die Kantenindikatormatrix Soziomatrix. Falls gleichzeitig nach Wahlen und Ablehnungen gefragt wird, werden in der Soziomatrix die Elemente gleich 1, 0 oder -1 gesetzt, je nachdem ob eine Wahl, weder Wahl noch Ablehnung oder Ablehnung vorliegt. Das zugehörige Modell ist der signierte Graph, der im nächsten Abschnitt definiert wird; wir werden hier jedoch die aus Wahlen bzw. Ablehnungen entstandenen gerichteten Graphen getrennt behandeln.

Ein Teil der soziometrischen Indizes beschreibt die Stellung der Individuen in dem sozialen System durch die Aus- und Eingangsgrade der zugehörigen Knoten des Graphen. Wenn die Kanten Wahlen bedeuten, ist der Ausgangsgrad a_i die Zahl der gegebenen Wahlen, der Eingangsgrad e_i die Zahl der erhaltenen Wahlen der Person i . Wenn die Anzahl der Wahlen nicht durch die Art der Untersuchung vorgegeben ist, kann bei einer Untersuchungsgruppe von n Personen jede Person höchstens $n-1$ Wahlen geben oder erhalten: die positive soziale Verbundenheit eines Individuums i wird nun durch $\frac{a_i}{n-1}$ gemessen, $\frac{e_i}{n-1}$ ist der Wahl-Status. Falls die

Kanten Ablehnungen bedeuten, heißen die entsprechenden Indizes negative soziale Verbundenheit und Ablehnungs-Status. Aus dem Verhältnis der gegebenen oder erhaltenen Wahlen können weitere Indizes gebildet werden, die hier nicht besprochen werden sollen. Interessant ist auch die Anzahl der gegenseitigen Wahlen, d.h. die Anzahl der Dyaden vom Typ M in Abb. 3.1.4. Formal können wir den ungerichteten Graphen betrachten, dessen Kanten wechselseitige Wahlen bedeuten und das positive Gegenseitigkeitsverhältnis wird durch $\frac{g_i}{n-1}$ gemessen, wobei g_i der Grad des Knotens i im ungerichteten Graphen, also die Anzahl der wechselseitigen Wahlen ist, an denen i teilhat.

Es gibt außerdem einige Kennzahlen, die die wechselseitigen und einseitigen Wahlen in Beziehung setzen; wir wollen jedoch zu Indizes übergehen, die die ganze Untersuchungsgruppe und nicht die einzelnen Individuen beschreiben. Dazu bieten sich die Durchschnittswerte der bisherigen Indizes für alle Individuen der Gruppe an; der Index der Interaktion in einer Gruppe ist die durchschnittliche soziale Verbundenheit. Nach Gl. (3.1.8) ist sie gleich dem durchschnittlichen Wahl-Status und beträgt $\frac{1}{n(n-1)}t_{y2}$, wobei t_{y2} die Größe des Graphen ist. In Abschnitt 3.1.7 hatten wir diese Größe Dichte genannt. Wenn wir den ungerichteten Graphen der wechselseitigen Wahlen betrachten, wird die Dichte dieses Graphen in der Soziometrie auch Index der Kohäsion genannt. Der Anteil der gegenseitigen Wahlen an allen gegebenen Wahlen heißt Index der Kohärenz.

Es muß angemerkt werden, daß in der Soziometrie oft die Anzahl der Wahlen pro Person durch die Frage festgelegt wird, wenn z.B. ein Schüler einen Sitznachbarn wählen soll oder eine Person nach ihren drei besten Freunden gefragt wird. In diesen Fällen sind die Ausgangsgrade, die Größe des gerichteten Graphen und damit auch die Mittel der Aus- und Eingangsgrade fest vorgegeben und liefern keine Information

über die Struktur der Gruppe. Von den genannten Indizes sind dann die der positiven (bzw. negativen) sozialen Verbundenheit und der Interaktion wertlos.

Am Schluß dieses Unterabschnittes zur Soziometrie sei darauf hingewiesen, daß hier nur ein sehr kleiner Teil von Indizes und soziometrischen Fragestellungen genannt wurde; andere Probleme der Soziometrie sind z.B. das Feststellen von Cliques, andere Indizes sind z.B. auf "triad counts" aufgebaut.

3.1.9 Bewertete Graphen

Anstatt nur anzugeben, ob eine Beziehung zwischen zwei Einheiten vorliegt oder nicht, ist es oft auch möglich und sinnvoll, die Stärke der Beziehung zu messen; also formal jeder Kante einen Wert zuzuordnen. Falls jeder Kante eine reelle Zahl zugeordnet ist, heißt der Graph bewertet oder Netzwerk. Man unterscheidet zwischen gerichteten und ungerichteten bewerteten Graphen, je nachdem, ob der zugrundeliegende Graph gerichtet oder ungerichtet ist.

In der Soziometrie kann man z.B. die Befragten bitten, die anderen Versuchspersonen je nach Sympathie in eine Punkteskala einzustufen oder man kann nach der Zahl der Besuche in einem gegebenen Zeitraum fragen. Wenn man nach der affektiven Beziehung fragt, ist es üblich, mehrere verschiedene Fragen dazu zu stellen; z.B.: "Von wem oder wem würden Sie Geld leihen, wen besuchen Sie oft, nennen Sie Ihre Freunde" usw.. Man kann nun versuchen, aus diesen Fragen eine Skala zu erstellen, die möglichst auf metrischem und nicht nur auf ordinalem Niveau die Stärke der Beziehung angibt. Wenn die Untersuchungseinheiten Unternehmen sind, kann man die Größe der Beteiligungen oder der Zahlungen eines Betriebes an den

anderen als Kantenwert darstellen; wenn die Kanten Verkehrswege zwischen Orten sind, kann man ihnen als Wert das Verkehrsaufkommen oder die Kapazität zuordnen. Bei einem Kommunikationsnetzwerk kann man den Kanten den Umfang der in einem bestimmten Zeitraum beobachteten Kommunikation zuordnen.

Bei einer zeichnerischen Darstellung eines bewerteten Graphen werden an den Verbindungslinien bzw. Pfeilen die Werte eingezeichnet. Die Wertmatrix y eines bewerteten Graphen wird durch die Elemente y_{ij} mit

$$(3.1.11) \quad y_{ij} = \begin{cases} \text{Wert der Kante } (i,j) \text{ bei gerichteten, der} \\ \text{Kante } \{i,j\} \text{ bei ungerichteten Graphen} \\ 0 \text{ falls } (i,j) \notin M \text{ bzw. } \{i,j\} \notin M \end{cases}$$

definiert. Bei ungerichteten Graphen ist sie symmetrisch und wir werden ihre Elemente mit $\bar{y}_{ij}$ bezeichnen. Die Kantenindikatormatrizen kann man als spezielle Wertmatrizen auffassen; sie ordnen jeder Kante den Wert 1 zu. Falls jeder Kante die Werte +1 oder -1 zugeordnet werden, je nachdem, ob eine positive oder negative Beziehung zwischen den Knoten vorliegt, heißt der zugehörige bewertete Graph signierter Graph. Im Zusammenhang mit der Soziometrie haben wir signierte Graphen schon erwähnt, sie spielen außerdem eine große Rolle in der Theorie der strukturellen Balance.

Es sei angemerkt, daß man anhand der Wertmatrix nicht unterscheiden kann, ob einer Kante (i,j) der Wert 0 zugeordnet wird oder ob keine Kante (i,j) existiert. Am Beispiel des Verkehrs- oder Kommunikationsnetzes wäre nicht zu unterscheiden, ob keine direkte Verkehrs- oder Kommunikationsverbindung zwischen zwei Knoten besteht oder ob kein Verkehr bzw. keine Kommunikation zwischen beiden Knoten stattfindet. In den meisten Anwendungen und speziell in dieser Arbeit werden diese beiden Möglichkeiten nicht unterschieden.

Formal gehen wir davon aus, daß der zugrundeliegende Graph vollständig ist und somit jedem gerichteten bzw. ungerichteten Knotenpaar ein Wert zugeordnet wird, der im Falle keiner Beziehung 0 beträgt.

Parameter eines bewerteten Graphen definieren wir wie in Gl. (3.1.1) - (3.1.10), wobei die Größen y_{ij} bzw. $\bar{y}_{ij}$ nun den Wert der Kante angeben. Die Parameter t_{y2} , $t_{\bar{y}2}$ werden wir bei bewerteten Graphen Kantensumme nennen, die Bezeichnungen Grad, Ausgangs- und Eingangsgrad behalten wir auch bei bewerteten Graphen bei; sie werden in der Literatur auch Wert, Ausgangs- und Eingangswert eines Knotens genannt. Bei bewerteten Graphen ist es sinnvoll, auch die Streuung der Kantenwerte zu messen; als Varianz s_{y2}^2 definieren wir bei gerichteten Graphen:

$$(3.1.12) \quad s_{y2}^2 := \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n (x_{ij} - \frac{1}{n(n-1)} t_{y2})^2 ,$$

bei ungerichteten Graphen:

$$(3.1.13) \quad s_{\bar{y}2}^2 := \frac{2}{n(n-1)} \sum_{i=1}^n \sum_{j>i}^n (\bar{x}_{ij} - \frac{2}{n(n-1)} t_{\bar{y}2})^2 .$$

Im unbewerteten Graphen sind diese Varianzen schon durch t_{y2} gegeben, z.B. gilt bei gerichteten, unbewerteten Graphen:

$$(3.1.14) \quad s_{y2}^2 = \frac{1}{n(n-1)} t_{y2} (1 - \frac{1}{n(n-1)} t_{y2}) .$$

3.1.10 Andere Definitionen von Graphen, andere Arten von Netzwerken

Bei der Definition von Graphen werden oft Kanten zugelassen, die den gleichen Anfangs- und Endpunkt haben, also Kanten der Form (i,i) bzw. $\{i,i\}$ bei ungerichteten Graphen. Sie heißen Schlingen (loops) und wurden hier aus der Definition ausgeschlossen, da sie Eigenschaften der Knoten und nicht Beziehungen zwischen verschiedenen Knoten bezeichnen. Allerdings werden wir später im allgemeinen von Graphen mit Schlingen ausgehen (vgl. 3.3.1). Bei der Matrixdarstellung von Graphen mit Schlingen enthält die Diagonale auch Einsen, bei bewerteten Graphen Zahlen ungleich Null. Die hier definierten Graphen ohne Schlingen heißen auch einfach (simple).

Manchmal werden Mehrfachkanten zugelassen, d.h. zu einem Paar (i,j) von Knoten werden mehrere Kanten von i nach j unterschieden.

Bei der Beobachtung von sozialen Netzwerken werden normalerweise mehrere Merkmale erfaßt, die teilweise Untersuchungseinheiten (z.B. Alter, Geschlecht, sozialer Status von Personen), teilweise Beziehungen zwischen Untersuchungseinheiten (Freundschaften oder Kontakte von Personen, Weisungsbefugnis in einem bestimmten Rahmen über eine andere Person) zugeordnet sind. Wie im vorigen Unterabschnitt erläutert, wird oft ein Merkmal, wie Zuneigung, mittels mehrerer Fragen erhoben und zu jeder Frage kann man einen zugehörigen Graphen bilden. In diesen Fällen liegen zu einer Knotenmenge mehrere Variablen vor, die sich auf die Knoten beziehen und es liegen mehrere bewertete, unbewertete, gerichtete oder ungerichtete Graphen vor; sie können beschrieben werden durch Angabe mehrerer Wert- bzw. Kantenindikatormatrizen, wobei die Werte in der Diagonalen sich auf Eigenschaften der Knoten und die anderen Matrixelemente sich auf Eigenschaften der Kanten, d.h. Knotenpaare, beziehen. Ein solches Netzwerk heißt multiples Netzwerk; es wird gebraucht,

um die Wechselwirkungen zwischen verschiedenen Arten von Beziehungen und Eigenschaften der Knoten zu untersuchen.

3.1.11 Einblick in die Literatur zur Analyse sozialer Netzwerke

Aus der reichhaltigen Literatur zur Analyse sozialer Netzwerke sollen hier einige Arbeiten, hauptsächlich Übersichtsartikel, genannt werden und einige weitere Probleme angesprochen werden. Literatur zu stochastischen Netzwerken und zur Inferenz aus Netzwerken wird in den beiden nächsten Abschnitten angegeben, auch wenn sich die Gebiete überlappen und in einigen hier erwähnten Arbeiten auch stochastische Fragestellungen behandelt werden. Die Bedeutung der Netzwerkanalyse wird unterstrichen durch die Existenz einer speziellen Zeitschrift: "Social Networks" und einiger Programmpakete, z.B. GRADAP; in anderen Zeitschriften, wie z.B. "Journal of Mathematical Sociology", "Quality and Quantity", "Sociological Methods & Research" finden sich oft Artikel zur Analyse sozialer Netzwerke.

An deutschsprachigen Einführungen seien der Abschnitt von Nehnevajsa zur Soziometrie im Handbuch der empirischen Sozialforschung, das Kapitel zur Graphentheorie in "Logik und Mathematik für Sozialwissenschaftler" von Kliemann und Müller (1976) und ein Übersichtsartikel von Kähler (1975) genannt. Kliemann, Müller bieten im wesentlichen eine Einführung in die Graphentheorie mit sozialwissenschaftlichen Beispielen, wichtige Gebiete der Analyse sozialer Netzwerke, wie z.B. triad counts oder Blockmodelle, werden nicht behandelt; Kähler gibt eine Einführung in die Literatur, wobei die Einordnung der Netzwerkanalyse in die Sozialwissenschaften und Anwendungen gegenüber der mathematischen Beschreibung des Modells im Vordergrund stehen.

Grundlegende Werke zur Anwendung der Graphentheorie in den Sozialwissenschaften sind die Bücher von Harary, Norman, Cartwright (1965) und Flament (1963). Zu neueren Ergebnissen der Netzwerkanalyse seien die Bibliographie von Freeman (1976) und der Übersichtsartikel von Burt (1980)

"Models of network structure" genannt. Burt unterteilt das Gebiet der Netzwerkanalyse in 6 Typen nach Aggregations-ebene und danach, ob Positionen von Untersuchungseinheiten oder Beziehungsmuster (positional oder relational approach) untersucht werden sollen. In mathematisch einheitlicher Bezeichnungsweise stellt er die Netzwerkanalyse und ihre Anwendung in den Sozialwissenschaften dar und gibt viele Bezüge zur Literatur an. Weitere Überblicke geben die Artikel von Felling (1975), Sorensen/Sorensen (1977), Burt (1978), und Niemöller/Schijf (1980).

Zu den in diesem Abschnitt definierten Parametern sollen zwei Aufsätze angegeben werden: Friedkin (1981) untersucht mittels Simulation die Effekte des Ansteigens der Netzwerkdichte auf fünf andere Maße der Kohäsion von Netzwerken, die durch Betrachtung von Triaden oder Wegen zwischen Knoten gewonnen werden. Zu zusammenhängenden Knoten i und j ist ein Weg eine Kantenfolge, die i und j verbindet. Im Artikel von Snijders wird die Varianz der Grade (bzw. Aus- und Eingangsgrade) als Maß der Heterogenität von Graphen untersucht und Abschätzungen ihrer Größe angegeben.

Zur Bedeutung der "triad-" und "dyad counts" sei besonders auf den Aufsatz von Holland und Leinhardt (1975) hingewiesen.

Abschließend wollen wir noch die Gewinnung soziometrischer Daten bzw. empirischer Netzwerke betrachten. In soziometrischen Untersuchungen wird meistens nach den Freunden und Bekannten nicht (nur) direkt, sondern mittels anderer Formulierungen gefragt, wie z.B.: "Wem würden Sie Geld leihen", "wem würden sie im Urlaub Ihre Wohnungsschlüssel anvertrauen",

"mit wem würden Sie persönliche oder berufliche Probleme bereden" usw.. McAllister/Fischer (1978) diskutieren geeignete Fragestellungen dieser Art, Fischer (1982) untersucht später, wie die Bezeichnung Freund mit Antworten auf indirekte Fragen korreliert, wobei die Ergebnisse natürlich zunächst nur für das Sozialgefüge und den Sprachgebrauch in den USA zutreffen.

Der Unterschied zwischen dem, was Personen über ihre Beziehungen sagen und ihrem Verhalten (cognitive-behavioral relations) wird von Bernard/Killworth/Sailer (1980, 1981) und Burt/Bittner (1981) diskutiert, wobei Bernard et al. anhand einiger Fallstudien zum Schluß kommen: "We are now convinced that cognitive data about communication can not be used as a proxy for the equivalent behavioral data". Dieser Hinweis auf die Ungenauigkeit bzw. Unklarheit einiger sozialwissenschaftlicher Daten soll darauf aufmerksam machen, daß das statistische Problem möglichst genauer Schätzung (z.B. des Mittels dieser Daten) eine kleinere Rolle spielt als in der Technik, den Naturwissenschaften, der Demographie oder den Wirtschaftswissenschaften, wo man meistens genau weiß, was die Daten in welcher Weise messen.

Die hier angesprochenen Probleme der Operationalisierung der Frage nach sozialen Beziehungen sind als Hintergrund zu den folgenden statistischen Überlegungen relevant, sie sind aber nicht Thema dieser Arbeit. Deshalb werden wir im weiteren Verlauf oft von "der" Frage nach Bekannten oder Freunden sprechen, ohne auf konkrete Formulierungen eines Fragebogens einzugehen.

3.2 Zufallsgraphen

3.2.1 Der Begriff des Zufallsgraphen

Eine Menge von Graphen mit einem Wahrscheinlichkeitsmaß (auf einer σ -Algebra dieser Menge) heißt Zufallsgraph, bei bewerteten Zufallsgraphen spricht man auch von Zufallsnetzwerk. Die Realisierungen von Zufallsgraphen sind Graphen; man kann Graphen als spezielle Zufallsgraphen auffassen, indem man einer Realisierung das Wahrscheinlichkeitsmaß 1 zuordnet.

Wir werden hauptsächlich Zufallsgraphen betrachten, bei denen die Graphen der Grundmenge alle dieselbe Knotenmenge $\Omega = \{1, \dots, n\}$ haben. In diesem Fall ist das Wahrscheinlichkeitsmaß durch die gemeinsame Verteilung der Zufallsvariablen $Y_{ij}, i \in \Omega, j \in \Omega, i \neq j$ bestimmt, wobei Y_{ij} den Kanten (i, j) bzw. $\{i, j\}$ Werte oder bei unbewerteten Graphen 0 oder 1 zuzuordnen; bei ungerichteten Graphen liegt also der in Abschnitt 2.1 geschilderte Fall vor, daß Paaren von Untersuchungseinheiten Zufallsvariablen zugeordnet sind.

Bei Zufallsgraphen ist die Kantenindikatormatrix bzw. Wertmatrix eine mehrdimensionale Zufallsvariable; um sie von ihren Realisierungen zu unterscheiden, bezeichnen wir sie und ihre Elemente mit Großbuchstaben Y, Y_{ij} . Analog sind die in den Abschnitten 3.1.6, 3.1.7 und 3.1.9 definierten Parameter von Graphen nun Zufallsvariablen, die wir auch mit den entsprechenden Großbuchstaben, z.B. $T_{Y2}, G_i, S_g^2, E_i, A_i$, bezeichnen und wir nennen den Erwartungswert von T_{Y2} : $\theta_{Y2} = E(T_{Y2})$. Parameter des Zufallsgraphen sind nun Parameter der Verteilung dieser Zufallsvariablen, z.B. θ_{Y2} als Erwartungswert von T_{Y2} , $E(G_i)$, $\text{var}(G_i)$, $E(S_g^2)$ usw., oder auch Parameter der Verteilung von "triad counts".

3.2.2 Meßfehlermodelle bei bewerteten Graphen

Bei bewerteten Graphen liegt es oft nahe, davon auszugehen, daß die Messungen der Kantenwerte mit zufälligen Fehlern behaftet sind. Wenn man z.B. aufgrund einiger Fragen oder subjektiver Einschätzungen eine Skala für die Sympathie zwischen Personen hat, wird man davon ausgehen müssen, daß bei einigen Untersuchungseinheiten die Skala unangemessen den Grad der Sympathie wiedergibt, weil einige Fragen bzw. Einschätzungen von Befragten untypisch beantwortet wurden, weil Fragen nicht richtig verstanden wurden, weil vom Interviewer oder bei der Dateneingabe Fehler gemacht wurden. Diese Abweichungen kann man nun als Zufallsfehler auffassen und aus inhaltlichen Gründen oder mit Hilfe von Erfahrungswerten wird man versuchen, Annahmen an die Verteilung der Fehler zu stellen.

Ein anderes Beispiel für Meßfehler liegt vor, wenn die Kantenwerte nicht beobachtet, sondern geschätzt werden. Wenn man den Autoverkehr zwischen Städten untersucht, ist es normalerweise unmöglich anzugeben, wieviel Autos in einem bestimmten Zeitraum von Stadt i nach Stadt j fahren; man kann nur aufgrund stichprobenartiger Verkehrszählungen oder mittels anderer Methoden Schätzungen für das Verkehrsaufkommen zwischen diesen beiden Städten erhalten.

Die beobachteten Variablen im Modell der zufälligen Meßfehler werden in der Form dargestellt:

$$(3.2.1) \quad Y_{ij} = \mu_{ij} + \varepsilon_{ij} \quad ,$$

wobei μ_{ij} der "wahre" Wert und ε_{ij} der Meßfehler ist. Normalerweise nimmt man an, daß die ε_{ij} unabhängig voneinander und identisch verteilt mit Erwartungswert 0 sind; bei dem oben erwähnten Beispiel eines Verkehrsnetzes wird man allerdings annehmen müssen, daß die Varianzen der Schätz-

funktionen und damit die Varianzen der ε_{ij} verschieden sind. Man wird diese Varianzen gegebenenfalls schätzen müssen. Die anderen beiden Annahmen bedeuten, daß ein Meßfehler bei Beobachtung eines Kantenwertes nicht den Meßfehler bei einer anderen Kante beeinflusst und daß die Meßfehler sich im Mittel ausgleichen. Wenn man z.B. nach der Anzahl der Kontakte zwischen zwei Personen fragt, und man davon ausgeht, daß gelegentlich Kontakte vergessen werden aber keine dazuerfunden werden, trifft die Annahme $E(\varepsilon_{ij}) = 0$ nicht zu; man kann dann aber in Gl. (3.2.1) eine weitere nicht zufällige Größe für den systematischen Fehler einführen. Diese Größe muß allerdings bekannt sein, wenn man aus den Beobachtungen der Y_{ij} auf die μ_{ij} schließen will. Bei der Frage nach den Kontakten muß man wissen, daß im Mittel nur z.B. 90 % der tatsächlichen Kontakte zwischen i und j genannt werden.

3.2.3 Ein Meßfehlermodell bei unbewerteten Graphen

Bei unbewerteten Graphen beschreibt Frank (1981, 1979b) ein Meßfehlermodell, indem er den beobachteten, empirischen Graphen als einen Zufallsgraphen auffaßt, der durch zufälliges Entfernen oder Hinzufügen von Kanten aus einem unbekannten, zugrundeliegenden Graphen entsteht. Die Verteilung des Zufallsgraphen ist dann durch den zugrundeliegenden Graphen und die Übergangswahrscheinlichkeiten für das Hinzufügen bzw. Entfernen von Kanten bestimmt. Dieses Modell können wir auch durch Gl. (3.2.1) darstellen, wenn μ_{ij} gleich dem entsprechenden Wert der Kantenindikatormatrix des zugrundeliegenden Graphen ist und die Zufallsvariablen ε_{ij} für $\mu_{ij} = 1$ die Werte 0 und -1, für $\mu_{ij} = 0$ die Werte 0 und 1 annehmen können. Die stochastische Unabhängigkeit der ε_{ij} und damit der Y_{ij} dürfte in der Praxis oft gewährleistet sein; wenn man z.B. nach Freunden oder Bekannten

fragt, dürften die falschen Angaben der Person i nicht die falschen Angaben der Person j beeinflussen. Die Annahme der identischen Verteilungen ist allerdings problematischer; abgesehen davon, daß der Wertebereich der ε_{ij} von μ_{ij} abhängt, muß man oft auch von unterschiedlichen Wahrscheinlichkeiten $P(\varepsilon_{ij}=0)$ und damit von unterschiedlichen Varianzen ausgehen. Bei einer Umfrage können Personen unterschiedlich vertrauenswürdig sein und vor allem können z.B. Kontakte zu schlecht angesehenen Personen überdurchschnittlich häufig verschwiegen werden.

Das Hauptproblem bei diesem Fehlermodell besteht darin, daß es nur Sinn gibt, falls $E(\varepsilon_{ij}) = P(\varepsilon_{ij} \neq 0) \neq 0$ gilt und wir deshalb $E(\varepsilon_{ij})$ kennen müssen, um aus Y_{ij} Information über

μ_{ij} zu erhalten. Bei der Schätzung von $\sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \mu_{ij}$ muß man

zumindest eine Annahme über $E \left(\sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \varepsilon_{ij} \right)$ machen können.

3.2.4 Zur Auffassung der Kontakte als Realisierung von Zufallsvariablen

Die beobachteten Kontakte selbst kann man als Realisierungen von Zufallsvariablen auffassen, wenn man z.B. in einem bestimmten Zeitraum Kontakte zwischen Personen beobachtet oder nach diesen Kontakten fragt. Es kann nun sein, daß Person i die Person j gerade in dem Untersuchungszeitraum überdurchschnittlich viel besucht hat und daß der Forscher weniger an der Zahl der Besuche in dem gegebenen Zeitraum, als an der Durchschnittszahl der Besuche z.B. pro Monat interessiert ist; er wird also den beobachteten Wert y_{ij} als Realisierung der Zufallsvariable Y_{ij} "Zahl der Besuche der Person i bei Person j pro Monat" auffassen.

Die Bestimmung einer Wahrscheinlichkeitsverteilung oder einer Klasse von in Frage kommenden Wahrscheinlichkeitsverteilungen für den Zufallsgraphen wird allerdings bei diesem Modell recht kompliziert, man wird davon ausgehen müssen, daß die Verteilungen der Y_{ij} voneinander abhängen und daß die Zufallsvariablen Y_{ij} voneinander stochastisch abhängen.

Literatur zu meistens stark vereinfachten Modellen solcher Kontaktnetzwerke wird in Abschnitt 3.2.6 diskutiert, wobei diese vereinfachten Modelle oft die Rolle eines Gedankenexperiments spielen (z.B. bei der Frage, wie sich Zufallsgraphen verhalten, wenn Kontakte unabhängig voneinander auftreten) oder die Formulierung einer Nullhypothese bei Tests darstellen.

3.2.5 Zur Auffassung der Y_{ij} als Schätzfunktionen

Im Zusammenhang mit dem Meßfehlermodell in 3.2.2 wurde als Beispiel genannt, daß Y_{ij} Schätzfunktionen eines unbekannten Parameters μ_{ij} sind. Wir wollen nun auf solche Modelle näher eingehen, bei denen zunächst mehrere Realisierungen eines Zufallsgraphen beobachtet werden. Außer dem schon erwähnten Fall der Beobachtung von Autoverkehr, liegt ein weiteres Beispiel hierfür vor, wenn Personen mehrere Male zu verschiedenen Zeitpunkten nach der Anzahl ihrer Kontakte zu anderen Personen in der vorhergehenden Woche befragt werden.

Formal liegen Realisierungen von identisch verteilten Zufallsvariablen $Y_{ij1}, \dots, Y_{ij\kappa}$ vor mit $E(Y_{ijv}) = \mu_{ij}$ für alle $i, j \in \Omega$. Anstelle der κ verschiedenen Zufallsgraphen $Y_1, \dots, Y_\kappa$ betrachten wir nun den Zufallsgraphen, der durch erwartungstreue Schätzfunktionen $Y_{ij} = f_{ij}(Y_1, \dots, Y_\kappa)$

für μ_{ij} definiert wird, z.B. durch $Y_{ij} = \frac{1}{K} \sum_{v=1}^n Y_{ijv}$ bei unabhängigen Beobachtungen. Für den Fall von mehreren stochastisch unabhängigen Beobachtungen eines Zufallsgraphen wurde dieses Modell in 2.4 behandelt und in Gl. (2.4.10) - (2.4.12) wurden Schätzfunktionen Y_{ij} für μ_{ij} , für $\text{var}(Y_{ij})$ und $\text{cov}(Y_{ij}, Y_{\ell m})$ angegeben.

Falls die Beobachtungen nicht unabhängig sind oder falls die Verteilung des Zufallsgraphen sich bei den verschiedenen Beobachtungen ändern kann, gibt die Schätztheorie in vielen Fällen auch Möglichkeiten zur Ermittlung der Schätzfunktionen Y_{ij} , deren Varianzen und Kovarianzen. Die Angabe oder Herleitung solcher Schätzfunktionen ist allerdings nicht Thema dieser Arbeit. Die Ausnutzung von Struktureigenschaften des zugrundeliegenden Zufallsgraphen kann außerdem zu Schätzfunktionen Y_{ij} führen, die auch Funktionen von gewissen $Y_{\ell mv}$ mit $(i,j) \neq (\ell,m)$ sind. Da die späteren Ausführungen zu Stichprobengraphen im allgemeinen auf solche Zufallsvariablen Y_{ij} nicht anwendbar sind (vgl. 3.3.4), gehen wir darauf auch nicht weiter ein.

Es muß noch angemerkt werden, daß bei zugrundeliegenden unbewerteten Graphen durch die Schätzfunktionen Y_{ij} aus den beobachteten Realisierungen $y_1, \dots, y_K$ des unbewerteten Graphen im allgemeinen ein bewerteter Graph konstruiert wird.

3.2.6 Literatur zur Verteilung von Zufallsgraphen

Modelle, die nicht auf sozialwissenschaftliche Anwendungen ausgerichtet sind, sollen hier nicht besprochen werden. Außerdem fehlen Gebiete wie z.B. stochastische Präferenzgraphen oder stochastische Netzwerkprozesse und die zahlreichen Arbeiten von Frank werden zusammen in Abschnitt

3.3.8 besprochen, da sie hauptsächlich Stichproben aus Graphen behandeln. Holland und Leinhardt (1981) haben in einer sehr beachteten Arbeit eine Exponentialfamilie von Verteilungen für gerichtete unbewertete Zufallsgraphen angegeben. Sie setzen voraus, daß die Variablen $D_{ij} = (Y_{ij}, Y_{ji})$ stochastisch voneinander unabhängig sind und erhalten eine Exponentialfamilie mit der Größe des Graphen, den Aus- und Eingangsgraden und der Anzahl der wechselseitigen Kanten $\sum_{i=1}^n \sum_{j=1}^n Y_{ij} Y_{ji}$ als suffizienten Statistiken. Die Parameter der Verteilungen werden interpretiert, mögliche Hypothesen für Tests genannt und das Modell wird auf einen empirischen Datensatz angewandt. In der schon erwähnten Arbeit zu "triad-counts" geben Holland und Leinhardt für beliebige Zufallsgraphen Formeln für Erwartungswerte, Varianzen und Kovarianzen der "triad counts" an und untersuchen speziell Zufallsgraphen mit der U/MAN-Verteilung: Bei vorgegebener Knotenmenge und zu fest vorgegebenen "dyad counts" (vgl. Abschnitt 3.1.7, Abb. 3.1.4) ordnet diese Verteilung jedem Graphen die gleiche Wahrscheinlichkeit zu. Unter Gültigkeit der U/MAN-Verteilung geben Holland und Leinhardt eine näherungsweise standardnormalverteilte Testgröße zur Prüfung linearer Hypothesen über die "triad counts" an. Wasserman (1977) setzt diese Untersuchungen fort, indem er auch andere Verteilungen behandelt, z.B. Gleichverteilung bei gegebenen Ausgangsgraden.

Bei seiner Untersuchung der Gradvarianz als Kennzahl für Graphen behandelt Snijders (1981) auch die entsprechende Zufallsvariable bei Zufallsgraphen; er gibt Erwartungswerte von S_g^2 , S_a^2 , S_e^2 (vgl. Gl. (3.1.4)(3.1.9)(3.1.10)) unter bestimmten Verteilungsannahmen, z.B. der U/MAN-Verteilung, an und standardisiert die Zufallsvariablen.

Bei dem in 3.1.11 genannten Überblicksartikel geht Burt (1980) auf das Testen von Hypothesen ein; auch Nehnevajsa (1973) beschreibt Zufallsmodelle in der Soziometrie.

3.2.7 Erzeugung von Zufallsgraphen durch Stichprobenverfahren

Bisher haben wir Zufallsgraphen besprochen, die durch stochastische Meßfehlermodelle oder aus inhaltlichen Gründen (Abschnitt 3.2.4) erzeugt wurden. Falls die Daten der Grundgesamtheit durch einen festen Graphen beschrieben werden können, beobachtet man oft nur einen mit einer Zufallsstichprobe ausgewählten Teil dieses Graphen; der beobachtete Graph ist also die Realisierung eines Zufallsgraphen, dessen Verteilung durch das Stichprobenverfahren und den zugrundeliegenden Graphen bestimmt ist. Im nächsten Abschnitt werden wir Stichproben aus Graphen genauer beschreiben und die Schätzung von Parametern aus Stichprobengraphen sind der Kern dieses Kapitels. Es sei angemerkt, daß bei unbewerteten Graphen Stichproben aus Graphen durch das in Abschnitt 3.2.3 beschriebene Meßfehlermodell beschrieben werden können: das Stichprobenverfahren entfernt mit bestimmter Wahrscheinlichkeit Kanten des zugrundeliegenden Graphen ohne Kanten hinzuzufügen; dieses Modell werden wir allerdings nicht zur Beschreibung von Stichprobengraphen benutzen.

3.3 Stichproben aus Graphen, Problemstellung

3.3.1 Beschreibung der Daten der Grundgesamtheit

Gegeben sei eine Grundgesamtheit $\Omega = \{1, \dots, n\}$ von Untersuchungseinheiten und eine $n \times n$ Zufallsmatrix Y mit Elementen $Y_{ij}, i \in \Omega, j \in \Omega$. Die unbekannte Verteilung W der Zufallsmatrix gehöre zu einer Klasse $\mathcal{W}$ von n^2 dimensionalen Verteilungen. In der Sprache der Graphentheorie ist Ω die Knotenmenge, Y die Wert- bzw. Kantenindikatormatrix des Zufallsgraphen und W die unbekannte Verteilung des Zufallsgraphen.

Bei unbewerteten Graphen nehmen die Zufallsvariablen Y_{ij} nur die Werte 0 und 1 an, bei ungerichteten Graphen gilt $Y_{ij} = Y_{ji}$. Wie vorher schon gesagt, werden wir bei ungerichteten Graphen die Variablen mit $\bar{Y}_{ij}$ bezeichnen. Die Elemente Y_{ii} der Diagonalen sind Zufallsvariablen, die den Knoten zugeordnet sind, wir können sie als Werte von Schlingen eines Graphen deuten. Falls ein Graph ohne Schlingen bzw. ohne Knotenwerte betrachtet wird, setzen wir $Y_{ii} \equiv 0$. Es sei angemerkt, daß bei den Schätzproblemen die Variablen Y_{ii} und Y_{ij} mit $i \neq j$ stets getrennt behandelt werden; sie sind hier im Modell zusammengefaßt, um alle Daten mit einer Wahrscheinlichkeitsverteilung zu beschreiben.

In der Literatur und z.B. hier in Abschnitt 3.2.7 wird meistens der Fall behandelt, daß die Daten der Grundgesamtheit durch einen unbekannten, nicht zufälligen Graphen beschrieben werden; in dem nun genannten Modell bedeutet das, daß die Klasse $\mathcal{W}$ aus Verteilungen W besteht, die jeweils einer $n \times n$ Matrix y die Wahrscheinlichkeit 1 zuordnen. Zusammenfassend kann man sagen, daß ein vollständiger, gerichteter, bewerteter Zufallsgraph mit Schlingen zugrundegelegt wird und alle anderen Graphen als Spezialfälle betrachtet werden.

3.3.2 Notation der Parameter der Grundgesamtheit

Für ungerichtete Graphen entspricht die obige Beschreibung der Grundgesamtheit genau der von Abschnitt 2.1 mit den Zufallsvariablen $\bar{Y}_{ij}$ anstelle von $\bar{X}_{ij}$ bzw. $\bar{X}_{ii} = X_i$. Außerdem können wir die Ergebnisse von Kapitel I z.B. auf die Grade G_i anwenden und um zu bezeichnen, auf welche Zufallsvariablen Parameter der ersten beiden Kapitel sich bei der Anwendung auf Zufallsgraphen beziehen, fügen wir als ersten Index dieser Parameter eine Kennzeichnung der Zufallsvariablen hinzu. Bei Parametern, die sich auf die Wertmatrix $\bar{Y}$ beziehen, fügen wir ein $\bar{y}$ hinzu: z.B. entsprechen $a_{\bar{y}11}$, $a_{\bar{y}22}$, $\sigma_{\bar{y}2}^2$ den Parametern a_{11} , a_{22} , σ_2^2 ; bei Parametern, die sich auf die Grade G_i beziehen, fügen wir ein g hinzu, z.B. a_{g11} , σ_{g1}^2 . Speziell gilt für die Parameter $\theta_{\bar{y}1}$, $\theta_{\bar{y}2}$, θ_{g1} und $\theta_{\bar{y}3}$

$$(3.3.1) \quad \theta_{\bar{y}1} = E(T_{\bar{y}1}) = \sum_{i=1}^n E(\bar{Y}_{ii})$$

$$(3.3.2) \quad \begin{aligned} \theta_{\bar{y}2} &= E(T_{\bar{y}2}) = \sum_{i=1}^n \sum_{j>i}^n E(\bar{Y}_{ij}) \\ &= \frac{1}{2} \sum_{i=1}^n G_i = \frac{1}{2} \theta_{g1} \end{aligned}$$

$$(3.3.3) \quad \theta_{\bar{y}3} = \theta_{\bar{y}1} + \theta_{\bar{y}2}$$

Es sei angemerkt, daß die Parameter σ_u^2 , σ_{u1}^2 (Gl. (2.2.21), (2.2.45)) angewandt auf die Zufallsvariablen $\bar{Y}_{ij}$ gleich $\sigma_{g1}^2 = E(S_g^2)$ bzw. σ_{g11} sind.

Für gerichtete Graphen betrachten wir zunächst den durch

$$(3.3.4) \quad \bar{Y}_{ij} := Y_{ij} + Y_{ji}$$

definierten zugehörigen ungerichteten Graphen $\bar{Y}$ und bezeichnen für diesen ungerichteten Graphen die Parameter wie oben (vgl. Bem. (2.1.4)). Falls Parameter Eigenschaften Y_{ij} der gerichteten Tupel (i,j) bezeichnen, verwenden wir zur Kennzeichnung y statt $\bar{y}$; z.B.

$$(3.3.5) \quad \theta_{Y2} = E(T_{Y2}) = \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n E(Y_{ij}) = \theta_{\bar{Y}2}$$

$$(3.3.6) \quad \sigma_{Y2}^2 = E \left(\frac{1}{n(n-1)} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n (\bar{X}_{ij} - \frac{1}{n(n-1)} T_{Y2})^2 \right)$$

Die Parameter σ_{e1}^2 , σ_{e11}^2 , σ_{a1}^2 , σ_{a11}^2 sind wie σ_1^2 bzw. σ_{11}^2 (Gl. (1.3.15) (1.3.17)) definiert mit den Eingangsgraden E_i bzw. Ausgangsgraden A_i anstelle von X_i . Für θ_{Y2} gilt:

$$(3.3.7) \quad \theta_{Y2} = \theta_{\bar{Y}2} = \theta_{e1} = \theta_{a1}$$

Die Beziehung zwischen σ_{g1}^2 , σ_{e1}^2 und σ_{a1}^2 lautet:

$$(3.3.8) \quad \begin{aligned} \sigma_{g1}^2 &= E \left(\frac{1}{n} \sum_{i=1}^n (A_i + G_i - \frac{2}{n} T_{Y2})^2 \right) \\ &= \sigma_{a1}^2 + \sigma_{e1}^2 + 2 \sigma_{ae1} \end{aligned}$$

mit

$$(3.3.9) \quad \sigma_{ae1} := E \left(\frac{1}{n} \sum_{i=1}^n (A_i - \frac{1}{n} T_{Y2}) (E_i - \frac{1}{n} T_{Y2}) \right) .$$

Für nicht zufällige Werte in der Grundgesamtheit ist σ_{ae1} die Kovarianz der Aus- und Eingangsgrade und wir haben hier diesen Begriff auf zufällige Werte in der Grundgesamtheit verallgemeinert.

Für bewertete Graphen wurden die Parameter schon weitgehend in den ersten beiden Kapiteln erklärt und deshalb wird in diesem Kapitel das Schwergewicht auf der Erklärung der Parameter für unbewertete Graphen liegen.

3.3.3 Zur Auffassung der Werte als Realisierung von Zufallsvariablen

Die Auffassung der Werte als Realisierung von Zufallsvariablen ist allgemeiner als ihre Betrachtung als fest vorgegebene Zahlen, wie in den ersten beiden Kapiteln und Abschnitt 3.3.1 schon erklärt wurde. Die Behandlung der normalen Stichprobenprobleme mit festen, unbekannten Variablenwerten in der Grundgesamtheit ist außerdem einfach als Spezialfall zu behandeln; man hat nur alle auftauchenden Varianzen und Kovarianzen von Zufallsvariablen Y_{ij} gleich 0 zu setzen, wie es ebenfalls in Kapitel I und II durchgeführt wurde. Der Sinn in der hier vorgenommenen Verallgemeinerung ergibt sich aus den Überlegungen in Abschnitt 3.2; bei den Meßfehlermodellen hatten wir für alle Werte der Grundgesamtheit die Existenz von zufälligen Meßfehlern postuliert, in Abschnitt 3.2.4 war es aus inhaltlichen Gründen angebracht, die Kantenwerte (Besuche in einem bestimmten Monat) als Realisierung von Zufallsvariablen anzusehen und in Abschnitt 3.2.5 hatten wir den Fall besprochen, daß die beobachteten Werte y_{ij} Realisierungen von Schätzfunktionen sind. Normalerweise will man in diesen Fällen aus einer Stichprobe Parameter der Verteilung des Zufallsgraphen schätzen, z.B. $\theta_{y2} = E(T_{y2})$; in

Meßfehlermodellen bedeutet das, daß man nicht den Wert

einer Realisierung von $\sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n y_{ij}$ im ganzen Graphen

schätzen will, sondern $\sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \mu_{ij}$ (vgl. Gl. (3.2.1)), wo-

bei man allerdings $E \left(\sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \epsilon_{ij} \right)$ kennen muß. Bei der Frage

nach Besuchen (Abschnitt 3.2.4) ist man meistens nicht an der Zahl der Besuche innerhalb der Grundgesamtheit in einem bestimmten Monat, sondern an ihrem Durchschnitt, genauer Erwartungswert, interessiert. Falls man allerdings im letzten Beispiel doch an der Zahl der Besuche in einem bestimmten Monat interessiert ist, wird man als Modell nicht einen Zufallsgraph, sondern einen festen unbekannten Graph wählen.

Die Beschreibung der Grundgesamtheit als Zufallsgraphen bietet außerdem die Perspektive, bestimmte Verteilungsmodelle (vgl. die Literaturangaben in Abschnitt 3.2.6) mit Problemen der Stichprobentheorie zu verbinden.

3.3.4 Stichprobengraphen

Es wird nicht eine Realisierung y der ganzen Matrix Y sondern nur ein Teil von y beobachtet. Die Stichprobenindikatorvariablen, die anzeigen, welches Element der Matrix y beobachtet wird, bezeichnen wir mit S_{ij} :

$$(3.3.10) \quad S_{ij} = \begin{cases} 1 & \text{falls Element in } i\text{-ter Zeile und} \\ & j\text{-ter Spalte von } y \text{ beobachtet wird} \\ 0 & \text{sonst} \end{cases}$$

Das Stichprobenverfahren wird durch die gemeinsame Verteilung der $\{S_{ij}, i \in \Omega, j \in \Omega\}$ gegeben. In der Sprache der Graphentheorie zeigt S_{ij} an, ob zum Knotentupel (i, j) ein Kantenwert bzw. bei unbewerteten Graphen die Existenz oder Nichtexistenz einer Kante beobachtet wird. Die Variablen S_{ii} zeigen an, ob der Knoten i in der Stichprobe ist und sein Wert beobachtet wird. Wir werden diese Zufallsvariablen auch einfach indizieren, d.h. $S_i = S_{ii}$.

Bei der Anwendung der Ergebnisse aus den ersten beiden Kapiteln setzen wir allgemein voraus, daß die Zufallsmatrix Y und die Menge $\{S_{ij} : i \in \Omega, j \in \Omega\}$ stochastisch unabhängig sind.

Ein Beispiel für die Verletzung dieser Voraussetzung haben wir schon in 2.1 erwähnt: falls $\theta_{\bar{y}2}$ aus den Graden

$$G'_i = \sum_{\substack{j=1 \\ j \neq i}}^n \bar{Y}_{ij} S_{ij} \text{ des Stichprobengraphen geschätzt werden soll,}$$

sind im allgemeinen das Stichprobenverfahren und die beobachteten Zufallsvariablen G'_i abhängig voneinander.

Bei zugrundeliegenden festen Graphen, also bei fester, unbekannter Matrix y gilt die stochastische Unabhängigkeit stets. Allerdings gibt es Stichprobenverfahren, bei denen die Auswahl durch Eigenschaften des zugrundeliegenden Graphen beeinflusst wird. Ein Beispiel hierfür ist das "snowball sampling" von Goodman (1961), mit dem er Freundschaftsbeziehungen in einer Population untersucht hat. Ausgehend von einer Zufallsstichprobe von Knoten $S^{(0)} \subset \Omega$ werden auf der ersten Stufe alle Knoten $S^{(1)} \subset \Omega$ ausgewählt, die mit Knoten aus $S^{(0)}$ durch eine Kante verbunden sind; auf der zweiten Stufe werden alle Knoten $S^{(2)}$ ausgewählt, die mit Knoten aus $S^{(1)}$ verbunden sind, usw.. Falls im Extremfall im zugrundeliegenden Graph keine Knoten durch Kanten verbunden werden, gilt für die Stichprobenindikatorvariablen $S_{ij} = 1$, falls $i \in S^{(0)}$

und $j \in S^{(0)}$; falls im anderen Extremfall der zugrundeliegende Graph vollständig ist, gilt nach der ersten Stufe $S_{ij} = 1$ für $i \in \Omega, j \in \Omega$. Bei den hier behandelten Stichprobenverfahren ist stets die Auswahl unabhängig von dem zugrundeliegenden Graphen.

3.3.5 Symmetrische Stichproben von Knotenpaaren

Ein vom Konzept her einfaches Stichprobenverfahren ist die zufällige Auswahl von Knotenpaaren bzw. Elementen der Matrix Y mit einer symmetrischen Stichprobe aus der Menge aller Paare $\{(i,j): i \in \Omega, j \in \Omega, i \neq j\}$ bei gerichteten Graphen ohne Schlingen oder $\{(i,j): i \in \Omega, j \in \Omega, j > i\}$ bei ungerichteten Graphen ohne Schlingen. Ein Beispiel hierfür wäre die Schätzung des Autoverkehrs auf Autobahnen. Wenn Kreuzungen und Abzweigungen durch Knoten und die Straßenverbindungen durch Kanten dargestellt werden, kann man aus der Menge aller Kanten eine Stichprobe zufällig auswählen und zu diesen Kanten (d.h. Straßen) den Autoverkehr messen. In der soziologischen Anwendung ist allerdings eine symmetrische Stichprobe von Kantenpaaren meistens schlecht durchzuführen und unpraktisch; die im nächsten Abschnitt besprochenen Stichprobengraphen sind hier angebrachter.

3.3.6 Die Stichprobengraphen SI , SN , SV , SVN

Wenn man z.B. Freundschaftsbeziehungen untersucht, wird man aus der Grundgesamtheit eine Zufallsstichprobe von Personen wählen und diese nach ihren Freunden insgesamt oder in der Stichprobe befragen; in graphentheoretischer Sprache wählt man also eine Menge von Knoten aus und beobachtet zu dieser Menge nach vorher bestimmten Regeln eine Anzahl von Kanten. Frank (1971) behandelt in seinem Buch vier Verfahren, auf die wir in diesem Kapitel noch genauer eingehen werden. Mit einem symmetrischen Stichprobenverfahren wird eine Knotenstichprobe $S \subset \Omega$ zufällig

ausgewählt und bei gerichteten Graphen werden vier Beobachtungsregeln für die Kanten unterschieden:

SI: Die Werte aller Knotentupel innerhalb von S werden beobachtet.

SV: Die Werte aller Knotentupel (i,j) mit Anfangspunkt $i \in S$ werden beobachtet.

SN: Die Werte aller Knotentupel (i,j) mit Endpunkt $j \in S$ werden beobachtet.

SVN: Die Werte aller Knotentupel (i,j) mit Anfangs- oder Endpunkt aus S ($i \in S$ oder $j \in S$) werden beobachtet.

Die Stichprobengraphen bezeichnen wir wie die oben genannten Beobachtungsregeln mit SI, SV, SN, SVN. Für ungerichtete Graphen sind die Beobachtungsregeln SV und SN gegenstandslos, nur die Stichprobengraphen SI und SVN unterscheiden sich dann.

Den Stichprobengraph SI kann man als Subgraph des zugrundeliegenden Graphen auffassen und das Verfahren wird deshalb auch als "subgraph sampling" bezeichnet; das Verfahren SVN wird auch "star-sampling" genannt (Capobianco (1972), Frank (1981)). Es sei angemerkt, daß beim Stichprobengraphen SVN die nicht beobachteten Knoten und Kanten einen Subgraphen des zugrundeliegenden Graphen bilden. Falls den Knoten Werte zugeordnet sind, werden bei allen Stichprobengraphen nur die Werte der Knoten der Stichprobe S beobachtet.

In der folgenden Tabelle werden einige Eigenschaften der Stichprobengraphen zusammengefaßt, wobei vom allgemeinen Fall des vollständigen, gerichteten, bewerteten Graphen mit Schlingen ausgegangen wird und deshalb jedes Tupel (i,j) als Kante bezeichnet werden kann. Mit S_i ($= S_{ii}$) und S_j sind die Stichprobenindikatorvariablen der Knotenstichprobe gemeint, p_1 und p_2 sind gemäß Gl. (1.1.3) als

$P(i \in S) = E(S_i) = p_1$ und $P(i \in S \text{ und } j \in S) = E(S_i S_j) = p_2$
definiert und K ist der Umfang der Knotenstichprobe.

Art des Stichprobengraphen	SI	SV	SN	SVN
Beobachtung von Kanten zu Knotenstichprobe S	innerhalb S	von S	nach S	von und nach S
Knotenmenge des Stichprobengraphen	S	Ω	Ω	Ω
Knotenmenge, deren Werte beobachtet werden	S	S	S	S
Kantenmenge des Stichprobengraphen	$S \times S$	$S \times \Omega$	$\Omega \times S$	$(S \times \Omega) \cup (\Omega \times S)$
Struktur des Stichprobengraphen	Subgraph	Teilgraph	Teilgraph	Teilgraph
Darstellung von S_{ij} mittels S_i und S_j für $i \neq j$	$S_{ij} = S_i S_j$	$S_{ij} = S_i$	$S_{ij} = S_j$	$S_{ij} = S_i + S_j - S_i S_j$
$E(S_{ij})$ für $i \neq j$	p_2	p_1	p_1	$2p_1 - p_2$
Anzahl der Knoten, deren Werte beobachtet werden	K	K	K	K
Anzahl der beobachteten Kanten (i, j) mit $i \neq j$	$K(K-1)$	$K(n-1)$	$K(n-1)$	$K(2n-K-1)$

Abschließend sollen die Beobachtungsregeln am Beispiel der Beobachtung von Besuchen zwischen Personen veranschaulicht werden: aus einer Population wählt man eine Stichprobe S von Personen aus und beobachtet bei SI alle Besuche zwischen Personen der Stichprobe, bei SV alle Besuche, die Personen der Stichprobe in der ganzen Population gemacht haben, bei SN alle Besuche, die die Personen der Stichprobe von Personen der ganzen Population erhalten haben und bei SVN alle

Besuche der Personen der Stichprobe, die sie in der ganzen Population gemacht oder aus ihr erhalten haben. In der Praxis wird man die Besuche nicht beobachten, sondern die Personen der Stichprobe danach fragen. Auf ein Problem bei solchen Befragungen gehen wir im nächsten Abschnitt ein.

3.3.7 Andere Stichprobengraphen

Bei Umfragen und der Beobachtungsregel SI wird man eine Liste aller Personen der Stichprobe erstellen und die Personen nach ihren Freundschaftsbeziehungen zu den Personen dieser Liste befragen. Falls der Stichprobenumfang groß ist, kann man den Befragten oft nicht mehr zumuten, zu allen anderen Personen der Stichprobe Angaben zu machen. Die Arbeiten zum "network Sampling" (Granovetter (1976, 1977)), Morgan/Rytina (1977), Erickson/Nosanchuk/Lee (1981)) gehen davon aus, daß die Liste der im Fragebogen aufgeführten Personen mindestens 150 und höchstens 500 Personen umfassen sollte; bei Stichproben von mehr als 500 Personen sind also die oben besprochenen Stichprobenverfahren schwer durchzuführen. Granovetter wählt deshalb aus der Stichprobe $S^{(B)}$ der Befragten mit einer zweiten Stichprobe $S^{(L)}$ die Namen aus, die auf die Liste des Fragebogens kommen; Erickson u.a. wählen überhaupt getrennte Stichproben für die Befragten und für die Namensliste des Fragebogens.

Für ungerichtete Graphen behandelt Bloemena (1964) eine allgemeinere Beobachtungsregel: es werden zwei disjunkte Stichproben von Knoten $S^{(1)}$ und $S^{(2)}$ gezogen und dann alle Kanten innerhalb $S^{(1)}$, innerhalb $S^{(2)}$ und von $S^{(1)}$ nach $S^{(2)}$ getrennt betrachtet. Falls $S^{(2)} = \emptyset$, erhält man die Beobachtungsregel SI, falls $S^{(2)} = \Omega - S^{(1)}$ erhält man SVN, wobei die Kanten innerhalb von $S^{(2)}$ nicht beobachtet

werden. Bei den im Zusammenhang mit dem "network sampling" genannten Beobachtungsregeln kann man $S^{(1)} = S^{(B)} \cap S^{(L)}$ und $S^{(2)} = (S^{(B)} \cup S^{(L)}) - (S^{(B)} \cap S^{(L)})$ setzen, wobei wieder die (ungerichteten) Kanten innerhalb $S^{(2)}$ nicht beobachtet werden.

Bevor wir die Problemstellung dieser Arbeit präzisieren, soll ein Überblick über die Arbeiten von O. Frank gegeben werden.

3.3.8 Bemerkungen zu Arbeiten von Frank zur Inferenz aus Graphen

Zunächst seien zwei Übersichtsartikel von Frank (1980, 1981) zur Inferenz aus Graphen genannt, wobei insbesondere der neuere "A survey of statistical methods for graph analysis" sehr ausführlich mit einer umfangreichen Bibliographie ist.

Eine Grundlage für Inferenz aus Graphen bildet sein Buch "Statistical inference in graphs", in dem nach einer ausführlichen Einführung in die Graphentheorie und einer Erläuterung des Begriffs des Zufallsgraphen zunächst bei Stichprobengraphen SI, SV, SN und SVN die Summe der Kantenergebnisse, die Häufigkeitsverteilung der Grade und die Anzahl der Zusammenhangskomponenten geschätzt werden. Die Schätzung der Summe der Kantenergebnisse wird auch in zwei späteren Aufsätzen (1977 a,b) behandelt; da dieses Problem auch das Thema dieser Arbeit ist, werden wir in den folgenden Abschnitten noch genauer auf Franks Ergebnisse dazu eingehen. Weitere Ergebnisse zur Schätzung der Anzahl der Zusammenhangskomponenten gibt eine Arbeit von 1978.

Zu der symmetrischen Stichprobe der Knoten sei angemerkt, daß Frank den Begriff etwas weiter als in dieser Arbeit und als in der Literatur üblich faßt. In Satz 1.1.2 hatten wir gezeigt, daß ein symmetrisches Stichprobenverfahren durch die Größen $p_m, m=1, \dots, n$ bestimmt werden kann.

Frank legt nur p_1 , p_2 , p_3 und p_4 fest, ohne Bedingungen an die Existenz von $p_m, m > 4$ zu stellen. Für 5-elementige Mengen $\Omega'_5 \subset \Omega$ und $\Omega_5 \subset \Omega$ kann also gelten: $P(\Omega_5 \subset S) \neq P(\Omega'_5 \subset S)$. Außerdem sieht Frank bei der Behandlung von Stichproben aus Graphen den zugrundeliegenden Graphen immer als fest und nicht als Zufallsgraphen wie in dieser Arbeit an.

Weiterhin untersucht Frank in seinem Buch Flüsse in Netzwerken, gibt Parameter der Verteilung von stochastischen Kontaktgraphen an, bei denen die Kanten unabhängig voneinander identisch verteilt auftreten und behandelt mit stochastischen Präferenzgraphen statistische Vergleiche von Rangordnungen.

In dem Buch werden nicht Schätzungen von "dyad-", "triad"- oder allgemein "subgraph counts" behandelt. Ergebnisse hierzu finden sich in den Arbeiten von 1979.

3.3.9 Problemstellung, allgemeine Bemerkungen zu den dargestellten Anwendungsmöglichkeiten

Die Ausführungen dieses Kapitels dienten bisher außer zur Einführung von Begriffen dazu, einen Einblick in die Analyse sozialer Netzwerke zu geben. Sie geben somit eine Grundlage für die Einordnung der Ergebnisse dieser Arbeit in das Gebiet der Netzwerkanalyse. Im weiteren Verlauf soll geklärt werden, wie die Ergebnisse der ersten beiden Kapitel auf Probleme der Inferenz aus Graphen angewandt werden können. Wir verzichten hierbei auf eine ausführliche Wiederholung der Ergebnisse in der Sprache der Graphentheorie. In den letzten beiden Abschnitten diskutieren wir zusätzlich Probleme des "network sampling" durch Vergleiche von Stichprobengraphen.

Betrachtet werden lineare, symmetrische Schätzfunktionen für θ_{y1} , θ_{y2} bzw. $\theta_{\bar{y}2}$ der Form:

$$(3.3.11) \quad \hat{\theta}_{y1} = \beta(K) \sum_{i=1}^n Y_{ii} S_{ii}$$

$$(3.3.12) \quad \hat{\theta}_{y2} = \gamma(K_2) \sum_{i=1}^n \sum_{j \neq i}^n Y_{ij} S_{ij}$$

$$(3.3.13) \quad \hat{\theta}_{\bar{y}2} = \gamma(K_2) \sum_{i=1}^n \sum_{j \neq i}^n \bar{Y}_{ij} S_{ij}$$

mit $K = \sum_{i=1}^n S_{ii}$, $K_2 = \sum_{i=1}^n \sum_{j \neq i}^n S_{ij}$. K ist die Anzahl der beobachteten Knotenwerte, K_2 ist im gerichteten Fall die Anzahl der beobachteten Kantenwerte, im ungerichteten Fall das Doppelte dieser Anzahl.

Die θ_1^* und θ_2^* (vgl. Gl. (1.3.10), (2.2.15)) entsprechenden Schätzfunktionen sind durch $\beta(K) = \frac{n}{K}$ und $\gamma(K_2) = \frac{n(n-1)}{K_2}$, die Horvitz-Thompson-Schätzfunktionen durch Einsetzen von $E(K)$ bzw. $E(K_2)$ für K und K_2 gegeben.

Die Betrachtungen zur Zulässigkeit (Abschnitt 1.4, 2.3) dienen im weiteren hauptsächlich zur Begründung der Wahl von θ_{y1}^* , θ_{y2}^* und $\theta_{\bar{y}2}^*$. Außerdem zeigen diese Überlegungen, daß es bessere (allerdings nicht erwartungstreue) Schätzfunktionen als die von Frank ausschließlich gebrauchten Horvitz-Thompson-Schätzfunktionen gibt. Andere zulässige Schätzfunktionen behandeln wir nicht, weil das Problem

möglichst genauer Schätzung in den Sozialwissenschaften gegenüber anderen Problemen in den Hintergrund rückt (vgl. Abschnitt 3.1.11). Dem Verfasser erscheint es hier sinnvoller, sich auf gut interpretierbare, gleichartige Schätzfunktionen zu beschränken als jeweils verschiedene, unter bestimmten Modellvorstellungen optimale Schätzfunktionen zu wählen.

In Bemerkung 1.2.8 hatten wir allgemein festgestellt, daß die wesentlichen Eigenschaften von $\hat{\theta}$ als Schätzfunktion für θ ebenfalls für $\alpha\hat{\theta}$ als Schätzfunktion für $\alpha\theta$ ($\alpha \in \mathbb{R}$) gelten. Für das Risiko (bzw. die Varianz) gilt $E(\alpha\hat{\theta} - \alpha\theta)^2 = \alpha^2 E(\hat{\theta} - \theta)^2$. Auch wenn wir uns im weiteren hauptsächlich auf die Schätzung von $\theta_{\bar{y}2}$ usw. beschränken, erhalten wir damit gleichzeitig Aussagen über die Schätzung von konstanten Vielfachen von $\theta_{\bar{y}2}$. Speziell kann man die Ergebnisse ohne Probleme auf die Schätzung der sozialen Dichte, Kohäsion, Interaktion, der durchschnittlichen Größe des Freundes- bzw. Bekanntenkreises anwenden ($\alpha = \frac{1}{n}$, $\alpha = \frac{1}{n(n-1)}$, $\alpha = \frac{2}{n(n-1)}$, vgl. Abschnitte 3.1.6 - 3.1.8).

Weiterhin wird darauf eingegangen, welche anderen Parameter man bei verschiedenen Stichprobengraphen mit den Ergebnissen der ersten beiden Kapitel (insbesondere Abschnitte 1.5, 2.4 und 2.5) schätzen kann. Wir haben schon erklärt, daß z.B. Streuung der Grade ein geeignetes Maß für die Heterogenität eines Graphen oder die Anzahl der wechselseitigen Kanten eine in der Soziometrie bedeutende Kennzahl ist. Die Schätzung solcher Parameter ist also für den Anwender wichtig. Bei zugrundeliegenden Zufallsgraphen sei allgemein auf die Überlegungen am Ende der Abschnitte 1.5, 2.4, 2.5 hingewiesen: Information über die Varianzen und Kovarianzen der Y_{ij} erhält man nur durch mehrmalige Beobachtung des Stichprobengraphen. Die Schätzungen dieser

Varianzen und Kovarianzen benötigt man unter anderem zur Schätzung der Streuung von $E(G_i)$, einer Kennzahl, die z.B. in Meßfehlermodellen besser als $E(S_g^2) = \sigma_{g1}^2$ die Streuung der "wahren" Werte beschreibt.

Es sei noch auf den Anhang hingewiesen, in dem die Bezeichnungen der ersten beiden Kapitel denen dieses Kapitels gegenübergestellt werden.

3.4 Anwendung der Ergebnisse aus Kapitel I

3.4.1 Vorbemerkung

Das Besondere an der Inferenz aus Graphen besteht in der Berücksichtigung der Tatsache, daß nicht den Untersuchungseinheiten selbst, sondern Paaren von Untersuchungseinheiten Werte zugeordnet sind. Allerdings können einige Probleme der Inferenz aus Graphen mit der normalen Stichprobentheorie behandelt werden, wenn die interessierenden Eigenschaften als Eigenschaften von Untersuchungseinheiten aufgefaßt werden können und für die Elemente einer symmetrischen Stichprobe von Untersuchungseinheiten beobachtet werden. Da dieser Abschnitt Inferenz aus Graphen nur soweit behandelt, wie sie auf Schätzprobleme aus symmetrischen Stichproben zurückzuführen ist, werden wir uns entsprechend kurz fassen.

3.4.2 Schätzung von θ_{y1} und anderer Parameter der Verteilung der $Y_{ii}: i \in \Omega$

Bei den Stichprobengraphen, die durch eine Zufallsauswahl von Knoten und Beobachtungsregeln für die Kantenwerte entstehen, gehen wir von einem symmetrischen Stichprobenverfahren für die Knoten aus und legen fest, daß die Knotenwerte für genau die Knoten der Knotenstichprobe beobachtet werden (vgl. die Tabelle in Abschnitt 3.3.6 zu SI, SV, SN, SVN). Es liegt also das im ersten Kapitel beschriebene Problem mit $X_i = Y_{ii}$ vor und alle Aussagen gelten für die Schätzung von θ_{y1} , a_{y11} , a_{y12} , σ_{y1}^2 , σ_{y11}^2 , $\text{var}(T_{y1})$.

Für einen unbewerteten Graphen mit Schlingen nehmen die Variablen Y_{ii} nur die Werte 0 und 1 an und wegen $Y_{ii} = Y_{ii}^2$

gilt $\theta_{y1} = a_{y11}$. Deshalb sind die Parameter des Satzes 1.3.3 gut zu interpretieren: a_{y11} ist der Erwartungswert der Anzahl der Schlingen und $\frac{1}{2}a_{y12}$ ist der Erwartungswert der Schlingenpaare in dem Zufallsgraphen. In Satz 1.3.5 kann man $\sigma_1^2 = \sigma_{y1}^2$ durch

$$(3.4.1) \quad \sigma_{y1}^2 = \frac{1}{n} \theta_{y1} - \frac{1}{n^2} \theta_{y1}^2 - \frac{1}{n^2} \text{var}(T_{y1})$$

ersetzen.

Frank (1971, S. 96-105) beschreibt die Grundgesamtheit als einen nicht zufälligen Graphen und behandelt ausschließlich Horvitz-Thompson-Schätzfunktionen, $\tilde{\theta}_{y1}$ für θ_{y1} , $\tilde{a}_{y11}$ für a_{y11} und $\tilde{a}_{y12}$ für a_{y12} , deren Unzulässigkeit wir für $\tilde{\theta}_{y1}$, $\tilde{a}_{y11}$ in 1.4 und für $\tilde{a}_{y12}$ in 2.4 (mit $X_{ij}=Y_{ii}Y_{jj}$) gezeigt haben. Aus den Schätzfunktionen $\tilde{a}_{y11}$ und $\tilde{a}_{y12}$ bildet er analog zu unserem Vorgehen in 1.5 eine erwartungstreue Schätzfunktion für $\text{var}(\tilde{\theta}_{y1})$. Wir haben in Gl. (1.5.8) eine Schätzfunktion für $\text{var}(\theta_{y1}^*)$ angegeben, die aus den zulässigen, erwartungstreuen Schätzfunktionen a_{y11}^* und a_{y12}^* konstruiert wurde.

3.4.3 Schätzung von Parametern aus dem Stichprobengraphen SV

Wegen $S_{ij} = S_i$ und $K_2 = K(n-1)$ haben die linearen symmetrischen Schätzfunktionen für $\theta_{y2} = E\left(\sum_{i=1}^n A_i\right) = \theta_{a1}$ die Form (vgl. Gl. (3.3.12))

$$(3.4.2) \quad \hat{\theta}_{y2} = \gamma(K_2) \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n Y_{ij} S_{ij} = \beta(K) \sum_{i=1}^n A_i S_i = \hat{\theta}_{a1}$$

mit $\beta(K) = \gamma(K_2)$; bei der Schätzung von $\theta_{a1} = \theta_{y2}$ aus dem Stichprobengraphen SV und symmetrischer Knotenstichprobe erhalten wir wieder das Schätzproblem von Kapitel I mit $X_i = A_i$. Da die Zufallsvariablen A_i mehr Werte als 0 oder 1 annehmen können, sind die Parameter a_{a11} und a_{a12} nicht sehr gut interpretierbar und die Angabe der Varianz der Schätzfunktion in Satz 1.3.5 mittels der Parameter σ_{a1}^2 , $\text{var}(T_{a1})$ und θ_{a1}^2 ist angebracht. Diese Parameter können wiederum mit Hilfe der Ergebnisse aus 1.5 geschätzt werden; σ_{a1}^2 ist dabei als Erwartungswert der Varianz S_a^2 (Gl. 3.1.9) der Knotenausgangsgrade ein nützliches Maß für die Heterogenität der Knoten des Zufallsgraphen.

Wir wollen noch auf eine Konsequenz davon aufmerksam machen, daß wir hier die Parameter eines Zufallsgraphen nur mittels Linearkombinationen der beobachteten Ausgangsgrade schätzen. Bei einer Umfrage braucht man also nur nach der Zahl der Freunde, der Besuche usw. fragen und nicht danach, wen die befragten Personen besucht haben.

Der Hauptgrund für die Wahl anderer Stichprobengraphen besteht darin, daß oft die Ausgangsgrade schlecht oder sehr ungenau beobachtet werden können: z.B. bei der Frage nach der Anzahl der Bekannten muß man davon ausgehen, daß der Befragte die tatsächliche Anzahl nur sehr ungenau schätzt. Beim Vergleich der Stichprobengraphen SI und SVN und bei den Bemerkungen zu Granovetters "network sampling" kommen wir auf dieses Problem zurück.

Da der Stichprobengraph SV nur für gerichtete Graphen sinnvoll definiert ist, erübrigt sich hier die Betrachtung von ungerichteten Graphen. Frank (1971, S. 137) behandelt die Schätzung von θ_{y2} aus dem Stichprobengraphen SV nur am Rande für fest vorgegebene Stichprobengröße.

3.4.4 Schätzung von Parametern aus dem Stichprobengraphen SN

In diesem Fall gilt wegen $S_{ij} = S_j: \theta_{y2} = \theta_{e1}$, $\hat{\theta}_{y2} = \hat{\theta}_{e1}$ und der Unterschied zum Stichprobengraphen SV besteht nur darin, daß anstelle der Ausgangsgrade A_i die Eingangsgrade E_i betrachtet werden; deshalb brauchen wir den Stichprobengraphen SN nicht weiter zu besprechen.

3.4.5 Schätzung von Parametern aus symmetrischer Stichprobe von Knotenpaaren

Dieses Stichprobenverfahren betrachten wir nur für Graphen ohne Schlingen bzw. ohne Knotenwerte (vgl. 3.3.5). Wenn wir bei gerichteten Graphen die Menge aller Knotentupel $\{(i,j): i,j \in \Omega, j \neq i\}$ und bei ungerichteten Graphen die Menge der Knotenpaare $\{(i,j): i,j \in \Omega, j > i\}$ als Grundgesamtheit definieren, sind den Elementen dieser Grundgesamtheit die Kantenwerte Y_{ij} bzw. $\bar{Y}_{ij}$ zugeordnet. Wenn aus einer symmetrischen Stichprobe aus dieser Grundgesamtheit der Erwartungswert $\theta_{y2}(\theta_{\bar{y}2})$ der Summe der Kantenwerte geschätzt werden soll, unterscheidet sich das Schätzproblem nicht inhaltlich, sondern nur in der Bezeichnungsweise von Kapitel I, und man kann wieder alle Ergebnisse zur Schätzung von Parametern aus Kap. I anwenden. Allerdings ist in diesem speziellen Fall die Bezeichnungsweise zum Teil sogar widersprüchlich; hier bezeichnen wir mit Ω die Knotenmenge des zugrundeliegenden Graphen und in Kapitel I war Ω die Grundgesamtheit, aus der eine Stichprobe gezogen wurde, also bei Knotenmenge Ω und z.B. gerichtetem Graphen die Menge $\{(i,j): i,j \in \Omega, i \neq j\}$. Um zu klären, was man in Kapitel I mit der in Kapitel III gebrauchten Notation für Graphen (siehe insbesondere 3.3.1 und 3.3.2) einsetzen muß, stellen wir in der folgenden Tabelle beide Bezeichnungsweisen gegenüber.

Bezeichnung in Kapitel I	Bezeichnung in Kapitel III für	
	gerichtete Graphen	ungerichtete Graphen
-	$\Omega = \{1, 2, \dots, n\}$ Knotenmenge des Graphen	$\Omega = \{1, 2, \dots, n\}$ Knotenmenge des Graphen
Ω (Grundgesamt- heit)	$\{(i, j) : i, j \in \Omega, i \neq j\}$	$\{\{i, j\} : i, j \in \Omega, j > i\}$
n	$n(n-1)$	$\frac{1}{2}n(n-1)$
S_i	S_{ij}	S_{ij}
$K = \sum S_i$ (Anzahl der be- obachteten Wer- te)	$K_2 = \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n S_{ij}$	$\frac{1}{2}K_2 = \sum_{i=1}^n \sum_{j>i}^n S_{ij}$
X_i	Y_{ij}	$\bar{Y}_{ij}$
$T_1 = \sum_{i=1}^n X_i$	$T_{Y2} = \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n Y_{ij}$	$T_{\bar{Y}2} = \sum_{i=1}^n \sum_{j>i}^n Y_{ij}$
$\theta_1 = E(T_1)$	$\theta_{Y2} = E(T_{Y2})$	$\theta_{\bar{Y}2} = E(T_{\bar{Y}2})$
σ_1^2 Gl. (2.1.15)	σ_{Y2}^2 Gl. (3.3.6)	$\sigma_{\bar{Y}2}^2$ Gl. (2.2.20) mit $\bar{Y}_{ij}$ statt $\bar{X}_{ij}$
σ_{11}^2	σ_{Y21}^2 (*)	$\sigma_{\bar{Y}21}^2$
a_{11}	a_{Y22} (*)	$a_{\bar{Y}22}$
a_{12}	$a_{Y23} + a_{Y24}$ (*)	$a_{\bar{Y}23} + a_{\bar{Y}24}$
$\hat{\theta}_1, \theta_1^*, \tilde{\theta}_1$	$\hat{\theta}_{Y2}, \theta_{Y2}^*, \tilde{\theta}_{Y2}$	$\hat{\theta}_{\bar{Y}2}, \theta_{\bar{Y}2}^*, \tilde{\theta}_{\bar{Y}2}$

(*) nicht in dieser Arbeit explizit definiert

Die a_{11}^* , a_{12}^* usw. entsprechenden Schätzfunktionen sind komplizierter anzugeben, aber da beim Einsetzen der entsprechenden Bezeichnungen für n , K usw. keine weiteren Probleme auftauchen, lassen wir sie hier weg.

Bei unbewerteten Graphen sind die a_{11} und a_{12} entsprechenden Parameter gut zu interpretieren: wegen $Y_{ij} = Y_{ij}^2$ gilt

$$a_{\bar{Y}22} = \theta_{\bar{Y}2} \text{ bzw. } a_{Y22} = \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n E(Y_{ij}^2) = \theta_{Y2} \text{ und die } a_{12}$$

entsprechenden Parameter ($a_{Y23} + a_{Y24}$) bzw. ($a_{\bar{Y}23} + a_{\bar{Y}24}$) sind jeweils der Erwartungswert der Summe aller gerichteten bzw. ungerichteten Kantenpaare des Zufallsgraphen. Bei unbewerteten Graphen bietet sich also zur Varianzangabe der Schätzfunktion der Satz 1.3.3 an oder in Satz 1.3.5 das Ersetzen von $\sigma_1^2 = \sigma_{\bar{Y}2}^2$ (σ_{Y2}^2) durch (vgl. Gl. (3.1.14))

$$(3.4.3) \quad \sigma_{\bar{Y}2}^2 = \frac{2}{n(n-1)} \theta_{\bar{Y}2} - \frac{4}{n^2(n-1)^2} (\theta_{\bar{Y}2}^2 + \text{var}(T_{\bar{Y}2}))$$

bzw.

$$(3.4.4) \quad \sigma_{Y2}^2 = \frac{1}{n(n-1)} \theta_{Y2} - \frac{1}{n^2(n-1)^2} (\theta_{Y2}^2 + \text{var}(T_{Y2}))$$

3.5 Anwendung der Ergebnisse aus Kapitel II, der Stichprobengraph SI

3.5.1 Das Schätzproblem

Wegen $S_{ij} = S_i S_j$ liegt bei ungerichteten Graphen genau das Schätzproblem des Kapitels II mit $\bar{X}_{ij} = \bar{Y}_{ij}$ vor. Wie schon in Bemerkung 2.1.4 dargelegt wurde, können wir gerichtete Graphen behandeln, indem wir aus den Kantenwerten Y_{ij} des gerichteten Graphen durch $\bar{Y}_{ij} = Y_{ij} + Y_{ji}$ einen ungerichteten Graphen konstruieren. Wegen $K_2 = K(K-1)$ haben die linearen symmetrischen Schätzfunktionen für $\theta_{y2} = \theta_{\bar{y}2}$ die Form:

$$\begin{aligned} (3.5.1) \quad \hat{\theta}_{\bar{y}2} &= \gamma(K) \sum_{i=1}^n \sum_{j>i}^n \bar{Y}_{ij} S_i S_j \\ &= \gamma(K) \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n Y_{ij} S_i S_j = \hat{\theta}_{y2} \end{aligned}$$

und wir können uns auf die Betrachtung ungerichteter Graphen beschränken.

Bei bewerteten Graphen ist zur Angabe des Erwartungswertes der quadrierten Abweichung vom wahren Wert der Satz 2.2.5 angebracht mit den Parametern $\sigma_{\bar{y}2}^2$ als Maß der Streuung der Kantenwerte und $\sigma_u^2 = \sigma_{g1}^2$ als Maß der Streuung der Grade. Mit den Ergebnissen aus 2.4 können wieder die Parameter $\sigma_{\bar{y}2}^2$, σ_{g1}^2 usw. erwartungstreu geschätzt werden. Für unbewertete Graphen sind die $a_{\bar{y}22}$, $a_{\bar{y}23}$, $a_{\bar{y}24}$ gut zu interpretieren und wir werden in den nächsten Unterabschnitten darauf genauer eingehen.

Es sei noch einmal darauf hingewiesen, daß bei gerichteten Graphen nur der durch $\bar{Y}_{ij} = Y_{ij} + Y_{ji}$ erzeugte ungerichtete Graph betrachtet wird und wir deshalb z.B. aus Abschnitt 2.4 keine Schätzungen für σ_{Y2}^2 , σ_{e1}^2 , σ_{a1}^2 erhalten, da diese Parameter nicht durch $\bar{Y}_{ij}$ bestimmt sind.

Wir wollen klären, was für Information aus dem Stichprobengraphen SI bei Umfragen nach dem Freundes-Bekanntenkreis, nach Kontakten usw. benötigt wird. Für die Schätzung von $\theta_{\bar{Y}2}$ braucht man nur Grade des Stichprobengraphen zu kennen, da $\hat{\theta}_{\bar{Y}2} = \sum_{i=1}^n \frac{1}{2} \left(\sum_{\substack{j=1 \\ j \neq i}}^n \bar{Y}_{ij} S_j \right) S_i$ gilt, und der Statistiker

braucht von den befragten Personen nur die Information über die Anzahl der Freunde, der Kontakte innerhalb der Stichprobe. Wenn der Befragte eine Liste mit allen Personen der Stichprobe vorgelegt bekommt, braucht er seine Bekannten in der Stichprobe nicht zu nennen, sondern nur ihre Anzahl anzugeben. Wenn man allerdings andere Parameter, z.B. σ_{Y2}^2 , σ_{g1}^2 , schätzen will, muß man zunächst die Schätzfunktionen a_{Y21}^* , a_{Y23}^* , a_{Y24}^* berechnen können und dazu braucht man die Produkte $Y_{ij}Y_{\ell m}$ der Kantenwerte. In diesem Fall benötigt der Statistiker die Information, wen der Befragte (in der Stichprobe) besucht hat, mit wem er befreundet ist, usw..

3.5.2 Parameter bei unbewerteten, ungerichteten Graphen

Weil $\bar{Y}_{ij}$ nur 0 und 1 als Werte annehmen kann, liegt es hier näher, anstelle von Erwartungswerten, Varianzen und Kovarianzen mit Wahrscheinlichkeiten zu arbeiten, und es erscheint sinnvoll, sich auf Satz 2.2.1 zu beziehen. Wir definieren zunächst:

$$(3.5.2) \quad \bar{w}_{ij} := P(\bar{Y}_{ij}=1) = 1-P(\bar{Y}_{ij}=0) = E(\bar{Y}_{ij}) = E(\bar{Y}_{ij}^2)$$

$$(3.5.3) \quad \bar{w}_{ij\ell m} := P(\bar{Y}_{ij}=1 \text{ und } \bar{Y}_{\ell m}=1) = E(\bar{Y}_{ij} \cdot \bar{Y}_{\ell m})$$

Bei festen Werten betragen die $\bar{w}_{ij}$ und $\bar{w}_{ij\ell m}$ 0 oder 1 und es gilt $\bar{w}_{ij\ell m} = \bar{w}_{ij} \cdot \bar{w}_{\ell m}$. Bei Unabhängigkeit der Kantenwerte $\bar{Y}_{ij}$ und $\bar{Y}_{\ell m}$ gilt auch $\bar{w}_{ij\ell m} = \bar{w}_{ij} \cdot \bar{w}_{\ell m}$. Die Parameter $a_{\bar{Y}22}$, $a_{\bar{Y}23}$ und $a_{\bar{Y}24}$ lassen sich nun wie folgt schreiben (vgl. Gl. (2.2.3), (2.2.6), (2.2.7)):

$$(3.5.4) \quad a_{\bar{Y}22} = \sum_{i=1}^n \sum_{j>i}^n \bar{w}_{ij} = \theta_{\bar{Y}2}$$

$$(3.5.5) \quad a_{\bar{Y}23} = 2 \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \sum_{\substack{\ell>j \\ \ell \neq i}}^n \bar{w}_{ij\ell}$$

$$(3.5.6) \quad a_{\bar{Y}24} = 2 \sum_{i=1}^n \sum_{j>i}^n \sum_{\substack{\ell>i \\ \ell \neq j}}^n \sum_{\substack{m>\ell \\ m \neq j}}^n \bar{w}_{ij\ell m}$$

Nach Bemerkung 2.2.2 lassen sich diese Parameter auch gut interpretieren; $a_{\bar{Y}22}$ ist der Erwartungswert $\theta_{\bar{Y}2}$ der Anzahl der Kanten, $a_{\bar{Y}23}$ ist das Doppelte des Erwartungswertes der Anzahl aller Kantenpaare, die einen Knoten gemeinsam haben, und $a_{\bar{Y}24}$ ist das Doppelte des Erwartungswertes der Anzahl der Kantenpaare, die keinen Knoten gemeinsam haben.

Wegen $a_{\bar{y}22} = \theta_{\bar{y}2}$ gilt für die Streuung $\sigma_{\bar{y}2}^2$ (Gl. (2.2.20)).

$$(3.5.7) \quad \sigma_{\bar{y}2}^2 = \frac{2}{n(n-1)}\theta_{\bar{y}2} - \frac{4}{n^2(n-1)^2}\theta_{\bar{y}2}^2 - \frac{4}{n^2(n-1)^2}\text{var}(T_{\bar{y}2})$$

und daher kann man bei der Angabe von $E(\hat{\theta}_{\bar{y}2} - \theta_{\bar{y}2})^2$ mittels Satz 2.2.5 den $\sigma_{\bar{y}2}^2$ entsprechenden Parameter $\sigma_{\bar{y}2}^2$ durch eine Linearkombination von $\theta_{\bar{y}2}$, $\theta_{\bar{y}2}^2$ und $\text{var}(T_{\bar{y}2})$ ersetzen. Man spart also einen Parameter und erhält $E(\hat{\theta}_{\bar{y}2} - \theta_{\bar{y}2})^2$ aus den Parametern $\theta_{\bar{y}2} = E(T_{\bar{y}2})$, $\text{var}(T_{\bar{y}2})$ und $\sigma_{g1}^2 = \sigma_u^2$ als anschauliches Maß für die Streuung der Grade des zugrundeliegenden Zufallsgraphen.

3.5.3 Parameter bei unbewerteten, gerichteten Graphen

Bei gerichteten Graphen können die Variablen $\bar{Y}_{ij} = Y_{ij} + Y_{ji}$ die Werte 0, 1 oder 2 und die Produkte $\bar{Y}_{ij}\bar{Y}_{lm}$ die Werte 0, 1, 2 oder 4 annehmen. Daher ist es angebracht, zunächst von den Verteilungen der Y_{ij} auszugehen, wobei zu beachten ist, daß nun die Reihenfolge der Indizes i, j wichtig ist. Wir definieren:

$$(3.5.8) \quad w_{ij} := P(Y_{ij}=1) = 1 - P(Y_{ij}=0) = E(Y_{ij}) = E(Y_{ij}^2)$$

$$(3.5.9) \quad w_{ijlm} := P(Y_{ij}=Y_{lm}=1) = E(Y_{ij} \cdot Y_{lm})$$

Falls die Werte unabhängig voneinander sind, gilt $w_{ijlm} = w_{ij}w_{lm}$ und falls die Kantenwerte fest vorgegeben

sind, nehmen zusätzlich w_{ij} und w_{ijlm} nur 0 und 1 als Werte an. Für die w_{ijlm} gilt $w_{ijlm} = w_{lmij}$ und weiterhin gilt für Tupel $(i,j), (\ell,m)$:

$$(3.5.10) \quad E(\bar{Y}_{ij} \cdot \bar{Y}_{lm}) = w_{ijlm} + w_{ijml} + w_{jilm} + w_{jiml}.$$

Die Parameter $a_{\bar{Y}22}$, $a_{\bar{Y}23}$ und $a_{\bar{Y}24}$ des Satzes 2.2.1 lassen sich wie folgt schreiben (vgl. Frank (1977a) S. 85,86):

$$(3.5.11) \quad \begin{aligned} a_{\bar{Y}22} &= \sum_{i=1}^n \sum_{\substack{j>i \\ j \neq i}}^n E(Y_{ij} + Y_{ji})^2 \\ &= \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n (w_{ij} + w_{ijji}) = \theta_{Y2} + 2 \sum_{i=1}^n \sum_{j>i}^n w_{ijji} \end{aligned}$$

$$(3.5.12) \quad \begin{aligned} a_{\bar{Y}23} &= 2 \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \sum_{\substack{\ell>j \\ \ell \neq i}}^n E(Y_{ij} + Y_{ji})(Y_{i\ell} + Y_{\ell i}) \\ &= 2 \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \sum_{\substack{\ell>j \\ \ell \neq i}}^n (w_{ijil} + w_{ijli} + w_{jiil} + w_{jili}) \\ &= \sum_{(i,j,\ell,m) \in \Omega^4} w_{ijlm}, \end{aligned}$$

wobei über $\{(i,j,\ell,m) \in \Omega^4 : i \neq j, \ell \neq m, |\{i,j,\ell,m\}|=3\}$ summiert wird und

$$(3.5.13) \quad \begin{aligned} a_{\bar{Y}24} &= 2 \sum_{i=1}^n \sum_{j>i}^n \sum_{\substack{\ell>i \\ \ell \neq j}}^n \sum_{\substack{m>\ell \\ m \neq j}}^n E(Y_{ij} + Y_{ji})(Y_{\ell m} + Y_{m\ell}) \\ &= 2 \sum_{i=1}^n \sum_{j>i}^n \sum_{\substack{\ell>i \\ \ell \neq j}}^n \sum_{\substack{m>\ell \\ m \neq j}}^n (w_{ijlm} + w_{ijml} + w_{jilm} + w_{jiml}) \\ &= \sum_{(i,j,\ell,m) \in \Omega^4} w_{ijlm}, \end{aligned}$$

wobei über $\{(i,j,\ell,m) \in \Omega^4 : |\{i,j,\ell,m\}|=4\}$ summiert wird.

$a_{\bar{y}22}$ ist also der Erwartungswert der Anzahl der gerichteten

Kanten plus dem Erwartungswert der Anzahl $\sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n Y_{ij} Y_{ji}$

$= 2 \sum_{i=1}^n \sum_{j>i}^n Y_{ij} Y_{ji}$ der Doppelkanten, wobei darauf hingewie-

sen sei, daß hier Doppelkanten zweimal gezählt werden: als Doppelkante von i nach j und von j nach i .

Wir wenden uns noch einmal der Summe der Kanten zwischen i und j $\bar{Y}_{ij} = Y_{ij} + Y_{ji}$ zu. $\bar{Y}_{ij} \bar{Y}_{\ell m}$ ist immer die Anzahl der Kantenpaare, die aus je einer gerichteten Kante zwischen i und j und zwischen ℓ und m gebildet werden können. Daher ist nach Bemerkung 2.2.2 analog zum Fall des ungerichteten Graphen $a_{\bar{y}23}$ das Doppelte des Erwartungswertes der Anzahl der (gerichteten) Kantenpaare, die genau einen Knoten gemeinsam haben, und a_{y24} das Doppelte des Erwartungswertes der Anzahl der Kantenpaare, die keinen Knoten gemeinsam haben.

Bei der Darstellung von $E(\hat{\theta}_{y2} - \theta_{y2})^2$ durch Satz 2.2.5 ist im Gegensatz zu den im vorigen Unterabschnitt behandelten ungerichteten Graphen dieses Mal $\sigma_{\bar{y}2}^2$ nicht durch andere im Satz vorkommende Parameter zu ersetzen, sondern man braucht zusätzlich den Erwartungswert der Anzahl der Doppelkanten:

$$(3.5.14) \quad \sigma_{\bar{y}2}^2 = \frac{2}{n(n-1)} \theta_{y2} - \frac{4}{n^2(n-1)} \theta_{y2}^2 - \frac{4}{n^2(n-1)^2} \text{var}(T_{y2}) \\ + \frac{4}{n(n-1)} \sum_{i=1}^n \sum_{j>i}^n w_{ij} Y_{ij} Y_{ji} .$$

Da die Anzahl der Doppelkanten $2 \sum_{i=1}^n \sum_{j>i}^n Y_{ij} Y_{ji}$ ein in der

Graphentheorie und der Anwendung (vgl. 3.2.6) sehr wichtiger Parameter ist, ist die Ersetzung von $\sigma_{\bar{y}2}^2$ nach Gl.

(3.5.14) oft angebracht. Den σ_{u1}^2 entsprechenden Parameter σ_{g1}^2 kann man durch die Streuung σ_{a1}^2 der Ausgangsgrade, σ_{e1}^2 der Eingangsgrade und der "Kovarianz" σ_{ae1} nach Gl. (3.3.8) darstellen.

3.5.4 Behandlung von signierten Graphen

In der Soziometrie werden oft signierte Graphen betrachtet, bei denen $y_{ij} = 1$ ist, falls zwischen Personen i und j eine positive soziale Beziehung besteht, $y_{ij} = 0$ falls eine neutrale bzw. keine Beziehung zwischen den Personen besteht und $y_{ij} = -1$ gilt, falls eine negative Beziehung vorliegt. Wir wollen jetzt nur ungerichtete, signierte Zufallsgraphen betrachten und sie aus den einzelnen Graphen $\bar{Y}_p$ und $\bar{Y}_n$, die nur positive oder negative Beziehungen darstellen, konstruieren. Es sei:

$$(3.5.15) \quad \bar{Y}_{pij} = \begin{cases} 1 & \text{falls eine positive soziale Beziehung} \\ & \text{zwischen } i \text{ und } j \text{ gemessen wird} \\ 0 & \text{sonst} \end{cases}$$

$$(3.5.16) \quad \bar{Y}_{nij} = \begin{cases} 1 & \text{falls eine negative soziale Beziehung} \\ & \text{zwischen } i \text{ und } j \text{ gemessen wird} \\ 0 & \text{sonst} \end{cases}$$

Die Summe der jeweiligen Kantenwerte bezeichnen wir mit T_p , T_n ; ihre Erwartungswerte mit θ_p , θ_n . Den signierten Graphen definieren wir durch

$$(3.5.17) \quad \bar{Y}_{ij} = \bar{Y}_{pij} - \bar{Y}_{nij}$$

und es gilt: $T_{\bar{y}2} = T_p - T_n$, $\theta_{\bar{y}2} = \theta_p - \theta_n$. Solange wir nur die Anzahl der positiven oder negativen Beziehungen schätzen wollen, betrachten wir die zugehörigen unbewerteten Graphen; falls wir Aussagen über das Verhältnis der positiven und negativen Beziehungen erhalten wollen, können wir den signierten Graphen betrachten und die Differenz $\theta_{\bar{y}2} = \theta_p - \theta_n$ schätzen. Bei der Angabe von $E(\hat{\theta}_{\bar{y}2} - \theta_{\bar{y}2})^2$ gilt für die Parameter $a_{\bar{y}22}$ und $\sigma_{\bar{y}2}^2$:

$$(3.5.18) \quad a_{\bar{y}22} = E \left(\sum_{i=1}^n \sum_{j>i}^n \bar{y}_{ij}^2 \right) = \theta_p + \theta_n ,$$

$$(3.5.19) \quad \begin{aligned} \sigma_{\bar{y}2}^2 &= \frac{2}{n(n-1)} (\theta_p + \theta_n) - \frac{4}{n^2(n-1)^2} (\theta_p - \theta_n)^2 \\ &\quad - \frac{4}{n^2(n-1)^2} \text{var}(T_p - T_n) \\ &= \frac{2}{n(n-1)} \theta_p \left(1 - \frac{2}{n(n-1)} \theta_p \right) \\ &\quad + \frac{2}{n(n-1)} \theta_n \left(1 - \frac{2}{n(n-1)} \theta_n \right) \\ &\quad + \frac{8}{n^2(n-1)^2} \theta_p \theta_n + \frac{4}{n^2(n-1)^2} \text{var}(T_{\bar{y}2}) . \end{aligned}$$

In Satz 2.2.5 kann man also anstelle von $\sigma_{\bar{y}2}^2$ und $\theta_{\bar{y}2}^2$ die Parameter θ_p und θ_n einführen.

3.5.5. Simultane Schätzung von $\theta_{\bar{y}1}$ und $\theta_{\bar{y}2}$

Bei soziometrischen Untersuchungen hat man normalerweise den Personen Variablenwerte $\bar{y}_{ii}$ zugeordnet (z.B. das Alter, den sozialen Status, das Einkommen) und den Personenpaaren

Werte $\bar{y}_{ij}$ zugeordnet. Bei der simultanen Schätzung von $\theta_{\bar{y}1}$ und $\theta_{\bar{y}2}$ gibt Satz 2.5.1 die Kovarianz beider Schätzfunktionen an. Der Parameter $\sigma_{3\bar{y}g}$ (σ_{3xu} Gl. (2.5.3)) ist ein Maß für die Abhängigkeit der Knoten und Kantenwerte; seine Schätzung ist z.B. bei der Frage der Abhängigkeit der sozialen Kontakte vom Alter, Einkommen usw. interessant und ist aus dem Stichprobengraphen SI mit den Ergebnissen von 2.5 möglich.

Wenn z.B. Knoten Personengruppen (Personen eines Haushaltes, Schulklassen) repräsentieren, kann man Kontakte innerhalb der Gruppe als Knotenwerte und Kontakte zwischen Gruppen als Kantenwerte auffassen und man wird außer $\theta_{\bar{y}1}$ und $\theta_{\bar{y}2}$ auch die Gesamtzahl der Kontakte $\theta_{\bar{y}3} = \theta_{\bar{y}1} + \theta_{\bar{y}2}$ schätzen wollen. Mittels Gl. (2.5.18) kann man nun $E(\hat{\theta}_{\bar{y}3} - \theta_{\bar{y}3})^2$ aus $E(\hat{\theta}_{\bar{y}1} - \theta_{\bar{y}1})^2$, $E(\hat{\theta}_{\bar{y}2} - \theta_{\bar{y}2})^2$, den systematischen Fehlern $E(\hat{\theta}_{\bar{y}1} - \theta_{\bar{y}1})$ und $E(\hat{\theta}_{\bar{y}2} - \theta_{\bar{y}2})$ und aus $\text{cov}(\hat{\theta}_{\bar{y}1}, \hat{\theta}_{\bar{y}2})$ berechnen. $\hat{\theta}_{\bar{y}3}$ ist im allgemeinen keine symmetrische Schätzfunktion mehr, aber nach Bemerkung 2.5.4 ist es bei diesem Problem auch nicht sinnvoll, sich auf symmetrische Schätzfunktionen zu beschränken.

3.5.6 Bemerkung zu O. Franks Ergebnissen

Frank setzt wieder in seinem Buch (1971) und einer Arbeit (1977) feste Werte in der Grundgesamtheit voraus und behandelt ausschließlich Horvitz-Thompson-Schätzfunktionen. In seiner Arbeit (1977a) liefert er Ergebnisse zur Schätzung von $\theta_{y3} = \theta_{y1} + \theta_{y2}$ und gibt die Varianz durch Parameter an, die den hier gebrauchten Parametern a_{11} , a_{12} , a_{22} , a_{23} usw. entsprechen. Er stellt diese Parameter auf mehrere Arten dar, indem er die Summen umformt, und gibt diese Parameter als Funktion bestimmter "subgraph counts" an. Bei unbewerteten, ungerichteten Graphen stellt er die

Varianz der Horvitz-Thompson-Schätzfunktion als Funktion von $\theta_{\bar{y}3}$ und $\sum_{i=1}^n (g_i + \bar{y}_{ii})^2$ dar, bei unbewerteten, gerichteten Graphen braucht er zusätzlich die Anzahl der wechselseitigen Kanten $\sum_{i=1}^n \sum_{j=1}^n y_{ij} y_{ji}$, wobei hier die Schlingen auch als wechselseitige Kanten aufgefaßt werden. Für unbewertete Graphen ohne Schlingen haben wir diese Parameter in 3.5.2 und 3.5.3 diskutiert, wobei wir anstelle von $\sum_{i=1}^n g_i^2$ die Streuung σ_{g1}^2 betrachtet haben.

3.6 Anwendung der Ergebnisse aus Kapitel II, der Stichprobengraph SVN

3.6.1 Das Schätzproblem

Da für den Stichprobengraphen SVN S_{ij} durch $S_{ij} = S_i + S_j - S_i S_j$ definiert wird, gilt $K_2 = K(2n-K-1)$ und lineare, symmetrische Schätzfunktionen $\hat{\theta}_{Y3}$ für $\theta_{Y2} = \theta_{g1} - \theta_{\bar{Y}2}$ haben die Form:

$$\begin{aligned}
 (3.6.1) \quad \hat{\theta}_{Y3} &= \gamma(K) \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n Y_{ij} (S_i + S_j - S_i S_j) \\
 &= \gamma(K) \left[\sum_{i=1}^n A_i S_i + \sum_{j=1}^n E_j S_j - \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n Y_{ij} S_i S_j \right] \\
 &= \gamma(K) \sum_{i=1}^n G_i S_i - \gamma(K) \sum_{i=1}^n \sum_{j>i}^n \bar{Y}_{ij} S_i S_j \\
 &= \hat{\theta}_{g1} - \hat{\theta}_{\bar{Y}2}
 \end{aligned}$$

Wir können uns also wieder auf die Betrachtung von ungerichteten Graphen beschränken und brauchen nur Schätzfunktionen der Form $\hat{\theta}_{g1} - \hat{\theta}_{\bar{Y}2}$ betrachten, wobei die Koeffizienten $\gamma(K)$ für beide Schätzfunktionen $\hat{\theta}_{g1}$ und $\hat{\theta}_{\bar{Y}2}$ gleich sind. Es liegt das in Bemerkung 2.5.4 und 2.5.5 besprochene Problem der Schätzung einer Linearkombination von $\theta_1 = \theta_{g1}$ und $\theta_2 = \theta_{\bar{Y}2}$ durch eine Linearkombination $\hat{\theta}_{Y3}$ der Schätzfunktionen vor, wobei wegen $\frac{1}{2}\theta_{g1} = \theta_{\bar{Y}2}$ in diesem Fall erwartungstreue, lineare, symmetrische Schätzfunktionen $\hat{\theta}_{Y3}$ existieren.

Die Form der Schätzfunktion ist inhaltlich plausibel; bei dem Ausdruck $\sum_{i=1}^n G_i S_i$ werden die Kanten zwischen Knoten

der Stichprobe doppelt gezählt und deshalb subtrahiert man diese Anzahl $\sum_{i=1}^n \sum_{j>i}^n \bar{Y}_{ij} S_i S_j$ wieder. Aus diesem Grunde erscheint es auch nicht sinnvoll, diese beiden Zufallsvariablen verschieden zu gewichten, wie es im allgemeinen bei Schätzfunktionen der Form $\hat{\theta}_{g1} - \hat{\theta}_{\bar{y}2}$ möglich ist und in 2.5 behandelt wurde.

Es sei darauf hingewiesen, daß der Anwender außer Kantenwerten innerhalb der Stichprobe Realisierungen der Grade desselben zugrundeliegenden Zufallsgraphen beobachten muß, was bei manchen soziometrischen Fragestellungen nicht gewährleistet ist. Wenn man z.B. nach der Anzahl der Bekannten innerhalb einer Grundgesamtheit fragt und außerdem die Befragten auf einer Namensliste ihre Bekannten nennen und zählen sollen, wird man nicht das gleiche Ergebnis erwarten und daher nicht beide Variablen als Grad und Kantenwert desselben Zufallsgraphen auffassen können.

3.6.2 Die Berechnung von $E(\hat{\theta}_{\bar{y}3} - \theta_{\bar{y}2})^2$

In Bemerkung 2.5.5 wurde gesagt, daß im allgemeinen $E(\hat{\theta}_3 - \theta_3)^2$ von 9 Parametern abhängt. Da den X_i hier die Grade G_i und den $\bar{X}_{ij}$ die Kantenwerte $\bar{Y}_{ij}$ desselben Zufallsgraphen entsprechen, bedeuten einige dieser 9 Parameter dasselbe, die Formel für $E(\hat{\theta}_{\bar{y}3} - \theta_{\bar{y}2})^2$ vereinfacht sich und speziell gilt $T_1 = 2T_2 = 2T_{\bar{y}2} = T_{g1}$, $\sigma_1^2 = \sigma_{\bar{y}2}^2$,

$$(3.6.2) \quad \sigma_1^2 = \sigma_u^2 = \sigma_{3xu} = \sigma_{g1}^2 = E\left(\frac{1}{n} \sum_{i=1}^n \left(G_i - \frac{2}{n} T_{\bar{y}2}\right)^2\right),$$

$$(3.6.3) \quad \text{cov}(T_1, T_2) = \frac{1}{2} \text{var}(T_1) = 2 \text{var}(T_2) = 2 \text{var}(T_{\bar{Y}2})$$

$$= \frac{1}{2} \text{var}(T_{g1})$$

Nach Bem. 2.5.5 und den Sätzen 2.5.1, 2.2.5, 1.3.5 gilt nun:

$$(3.6.4) \quad E(\hat{\theta}_{\bar{Y}3} - \theta_{\bar{Y}2})^2 = \sigma_{\bar{Y}2}^2 h_1(\gamma) + \sigma_{g1}^2 h_2(\gamma) + \text{var}(T_{\bar{Y}2}) h_3(\gamma)$$

$$+ \theta_{\bar{Y}2}^2 h_4(\gamma)$$

mit:

$$(3.6.5) \quad h_1(\gamma) = f_1(\gamma) \quad (\text{Gl. (2.2.25)})$$

$$= E\left(\gamma^2(K) \frac{K(K-1)(n-K)(n-K-1)}{2(n-2)(n-3)}\right)$$

$$(3.6.6) \quad h_2(\gamma) = g_1(\gamma) + f_2(\gamma) \quad (\text{Gl. (1.4.2), (2.2.26)})$$

$$= E\left(\gamma^2(K) \frac{2K(K-1)(n-K)}{(n-1)(n-2)}\right)$$

$$= E\left(\gamma^2(K) \frac{K(n-K)(n-K-1)(n-K-2)}{(n-1)(n-2)(n-3)}\right)$$

$$(3.6.7) \quad h_3(\gamma) = 4g_2(\gamma) + f_3(\gamma) \quad (\text{Gl. (1.4.3), (2.2.27)})$$

$$= 4E\left(\gamma^2(K) \frac{K^2(K-1)}{n^2(n-1)}\right)$$

$$= E\left(\gamma^2(K) \frac{K^2}{n^2} \left(4 + \left(\frac{K-1}{n-1}\right)^2 - 4\frac{(K-1)}{(n-1)}\right)\right)$$

$$= E\left(\gamma(K) \frac{K(2n-K-1)}{n(n-1)}\right)^2$$

$$(3.6.8) \quad h_4(\gamma) = 4g_3(\gamma) + f_4(\gamma) \quad (\text{Gl. (1.4.4), (2.2.28)})$$

$$- 4E\left[\left(\gamma(K) \frac{K}{n} - 1\right) \left(\gamma(K) \frac{K(K-1)}{n(n-1)} - 1\right)\right]$$

$$= E\left(\gamma(K) \frac{K(2n-K-1)}{n(n-1)} - 1\right)^2$$

Für die durch $\gamma^*(K) = \frac{n(n-1)}{K(2n-K-1)}$ definierte Schätzfunktion $\theta_{\bar{Y}3}^*$ vereinfachen sich zwar die Ausdrücke $h_1(\gamma^*)$ und $h_2(\gamma^*)$ nicht wesentlich, aber es gilt $h_3(\gamma^*) = 1$ und $h_4(\gamma^*) = 0$ und $\theta_{\bar{Y}3}^*$ ist erwartungstreu:

$$\begin{aligned} (3.6.9) \quad E(\theta_{\bar{Y}3}^*) &= \sum_{i=1}^n E(G_i) E\left(\frac{n(n-1)}{K(2n-K-1)} S_i\right) \\ &\quad - \sum_{i=1}^n \sum_{j>i}^n E(\bar{Y}_{ij}) E\left(\frac{n(n-1)}{K(2n-K-1)} S_i S_j\right) \\ &= \theta_{\bar{Y}2} \left[E\left(\frac{2n(n-1)}{K(2n-K-1)} \frac{K}{n}\right) - E\left(\frac{n(n-1)}{K(2n-K-1)} \frac{K(K-1)}{n(n-1)}\right) \right] \\ &= \theta_{\bar{Y}2} \quad (\text{vgl. Gl. (1.1.28)}) \end{aligned}$$

O. Frank (1971) wählt statt $\theta_{\bar{Y}3}^*$ die entsprechende Horvitz-Thompson-Schätzfunktion und gibt die Varianz als Funktion der Parameter a_{22} , a_{23} , a_{24} an. Er braucht so einen Parameter mehr als bei Angabe der Varianz von $\theta_{\bar{Y}3}^*$ für feste Werte in der Grundgesamtheit, da man bei $\theta_{\bar{Y}3}^*$ nur die Parameter $\sigma_{\bar{Y}2}^2$ und σ_{g1}^2 benötigt.

Bei unbewerteten oder signierten Graphen gelten für die Parameter die Überlegungen aus 3.5.2 - 3.5.4 und man kann $\sigma_{\bar{y}2}^2$ gemäß Gl. (3.5.7), (3.5.14) oder (3.5.19) durch geeignete Parameter ersetzen.

3.6.3 Zur Schätzung anderer Parameter

Da man beim Stichprobengraphen Realisierungen der Grade für die Knoten der Stichprobe beobachtet, kann man die Streuung σ_{g1}^2 nach Gl. (1.5.5) erwartungstreu durch

$$(3.6.10) \quad \sigma_{g1}^{2*} = \frac{n-1}{n} \frac{1}{K-1} \sum_{i=1}^n \left(G_i - \frac{1}{K} \sum_{i=1}^n G_i S_i \right)^2 S_i$$

schätzen. Bei gerichteten Graphen erhält man analog Schätzfunktionen für die Streuungen der Aus- und Eingangsgrade σ_{a1}^2 und σ_{e1}^2 .

Andere Parameter wie z.B. $a_{\bar{y}22}$, $a_{\bar{y}23}$, $a_{\bar{y}24}$, $\sigma_{\bar{y}2}^2$ kann man mit den Ergebnissen von Abschnitt 2.4 erwartungstreu schätzen, wobei dann allerdings die größere Information, die der Stichprobengraph SVN gegenüber dem Stichprobengraphen SI bietet, nicht ausgeschöpft wird.

Schätzfunktionen, die die Information des Stichprobengraphen ausnutzen, kann man in gleicher Weise wie in Abschnitt 2.4 durch Betrachtung der Stichprobenindikatorvariablen $S_{ij} = S_i + S_j - S_i S_j$ anstelle von $S_i S_j$ erhalten. Wegen der großen Analogie werden hier nur als Beispiel erwartungstreue Schätzfunktionen für $a_{\bar{y}22} = \sum_{i=1}^n \sum_{j>i}^n (Y_{ij} + Y_{ji})^2$ und

$a_{y22} = \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n Y_{ij}^2$ bei gerichteten Graphen angegeben und das

weitere Vorgehen beschreiben.

Die Schätzfunktionen

$$(3.6.11) \quad a_{\bar{y}22}^* := \frac{n(n-1)}{K(2n-K-1)} \sum_{i=1}^n \sum_{j>i}^n (y_{ij} + y_{ji})^2 s_{ij}$$

$$(3.6.12) \quad a_{y22}^* := \frac{n(n-1)}{K(2n-K-1)} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n y_{ij}^2 s_{ij}$$

sind wegen der Unabhängigkeit der Werte vom Stichprobenverfahren und wegen

$$(3.6.13) \quad E\left(\frac{n(n-1)}{K(2n-K-1)} s_{ij}\right) = E\left(\frac{n(n-1)}{K(2n-K-1)} s_i\right) + E\left(\frac{n(n-1)}{K(2n-K-1)} s_j\right) - E\left(\frac{n(n-1)}{K(2n-K-1)} s_i s_j\right) = 1$$

(vgl. Gl. (1.1.28)) erwartungstreu. Etwas komplizierter, aber in ähnlicher Weise kann man nun erwartungstreue Schätzfunktionen für $a_{\bar{y}23}$, $a_{\bar{y}24}$ bzw. a_{y23} , a_{y24} bei gerichteten Graphen angeben und daraus Schätzfunktionen für $E(T_{y2}^2)$, $\sigma_{\bar{y}2}^2$ bzw. σ_{y2}^2 bestimmen.

Zusammenfassend sei gesagt, daß der Stichprobengraph SVN mehr relativ einfache Möglichkeiten zur Schätzung von Parametern bietet als die bisher behandelten Stichprobengraphen. Vergleiche der Varianz der Schätzfunktionen für θ_{y2} aus verschiedenen Stichprobengraphen werden in Abschnitt 3.7 behandelt.

3.7 Vergleich einiger Stichprobengraphen

3.7.1 Vorbemerkung

Wir wollen hauptsächlich die Stichprobengraphen SI und SVN vergleichen und können uns daher soweit auf ungerichtete Graphen beschränken. Am Schluß dieses Abschnittes wollen wir die Problematik des bei soziometrischen Fragestellungen wichtigen Vergleiches von SV und SI darstellen und auf Frank's Ergebnisse zum Vergleich von Stichprobengraphen hinweisen.

Da wir die Beobachtungsregeln vergleichen wollen, betrachten wir nicht Stichprobengraphen mit gleicher Anzahl

$K_2 = \sum_{i=1}^n \sum_{j>i}^n S_{ij}$ von beobachteten Kantenwerten, sondern Stichprobengraphen mit gleichem Auswahlverfahren für die Knoten. Der Sinn dieses Vergleichs soll an einer soziometrischen Fragestellung illustriert werden. Bei der Schätzung der Anzahl der Kontakte zwischen Personen einer Grundgesamtheit wird zunächst eine Stichprobe von Personen ausgewählt und zu diesen Personen werden Kontakte beobachtet. Der folgende Vergleich gibt dann Auskunft darüber, ob es sinnvoll ist, bei SI nur Kontakte zwischen Personen der Stichprobe oder zu jeder Person die Anzahl der Kontakte innerhalb der Grundgesamtheit oder bei SVN beides zu ermitteln.

Bei den Stichprobengraphen SI und SVN beschränken wir uns auf die Betrachtung der Schätzfunktionen:

$$(3.7.1) \quad \theta_{\bar{y}^2}^* = \frac{n(n-1)}{K(K-1)} \sum_{i=1}^n \sum_{j>1}^n \bar{y}_{ij} S_i S_j \quad \text{für SI und}$$

$$(3.7.2) \quad \theta_{\bar{y}3}^* = \frac{n(n-1)}{K(2n-K-1)} \sum_{i=1}^n \sum_{j>i}^n \bar{y}_{ij} (S_i + S_j - S_i S_j)$$

für SVN zu einer symmetrischen Knotenstichprobe mit Umfang K . Zusätzlich bietet sich noch der Vergleich von $\theta_{\bar{y}2}^*$ und $\theta_{\bar{y}3}^*$ mit einer anderen erwartungstreuen Schätzfunktion für $\theta_{\bar{y}2}$, die nur aus dem Mittel der beobachteten Grade gebildet wird:

$$(3.7.3) \quad \frac{1}{2}\theta_{g1}^* = \frac{n}{2K} \sum_{i=1}^n G_i S_i = \frac{n}{2K} \sum_{i=1}^n \sum_{j \neq i}^n \bar{y}_{ij} S_i$$

Die letzte Schätzfunktion ist zwar einfach und naheliegend, zählt allerdings die Kanten zwischen Knoten der Stichprobe im Gegensatz zu den vorher genannten Schätzfunktionen doppelt.

3.7.2 Die Varianzen von $\theta_{\bar{y}2}^*$, $\theta_{\bar{y}3}^*$ und $\frac{1}{2}\theta_{g1}^*$

In die Varianz von $\frac{1}{2}\theta_{g1}^*$ gehen nur die Parameter σ_{g1}^2 und $\text{var}(T_{\bar{y}2})$ ein, in die Varianzen der beiden anderen Schätzfunktionen zusätzlich $\sigma_{\bar{y}2}^2$. Um alle Varianzen zu vergleichen, setzen wir $\sigma_{\bar{y}2}^2 = \alpha \sigma_{g1}^2$ in der Grundgesamtheit voraus und variieren das Verhältnis α beider Streuungen. Für $\sigma_{g1}^2 = 0$ und $\sigma_{\bar{y}2}^2 > 0$ sieht man sofort aus den jeweiligen Formeln für die Varianz der Schätzfunktionen, daß θ_{g1}^* die beste der drei Schätzfunktionen ist. Allgemein gilt wegen

Gl. (2.2.23) : $\alpha \geq \frac{1}{(n-1)^2}$. Nach den Sätzen 1.3.5, 2.2.5

und Gl. (3.6.4) gelten für die Varianzen der Schätzfunktionen mit $\sigma_{\bar{y}2}^2 = \alpha \sigma_{g1}^2$.

$$(3.7.4) \quad \text{var}(\frac{1}{2}\theta_{g1}^*) = \sigma_{g1}^2 E(q_1(K)) + \text{var}(T_{\bar{y}2})$$

$$(3.7.5) \quad \text{var}(\theta_{\bar{y}2}^*) = \sigma_{g1}^2 E(q_2(K, \alpha)) + \text{var}(T_{\bar{y}2})$$

$$(3.7.6) \quad \text{var}(\theta_{\bar{y}3}^*) = \sigma_{g1}^2 E(q_3(K, \alpha)) + \text{var}(T_{\bar{y}2}) ,$$

wobei

$$(3.7.8) \quad q_1(K) := \frac{n^2(n-K)}{4K(n-1)}$$

$$\begin{aligned} (3.7.9) \quad q_2(K, \alpha) &:= \frac{\alpha n^2(n-1)^2(n-K)(n-K-1)}{2K(K-1)(n-2)(n-3)} \\ &+ \frac{n^2(n-1)(K-2)(n-K)}{K(K-1)(n-2)(n-3)} \\ &= \frac{n^2(n-1)(n-K)}{2K(K-1)(n-2)(n-3)} (\alpha(n-1)(n-K-1) + 2(K-2)) \end{aligned}$$

$$\begin{aligned} (3.7.10) \quad q_3(K, \alpha) &:= \frac{\alpha n^2(n-1)^2(K-1)(n-K)(n-K-1)}{2K(2n-K-1)^2(n-2)(n-3)} \\ &+ \frac{n^2(n-1)(n-K)(n-K-1)(n-K-2)}{K(2n-K-1)^2(n-2)(n-3)} \\ &= \frac{n^2(n-1)(n-K)(n-K-1)}{2K(2n-K-1)^2(n-2)(n-3)} (\alpha(n-1)(K-1) + 2(n-K-2)) \end{aligned}$$

Da die Koeffizienten von $\text{var}(T_{\bar{y}2})$ bei allen drei Schätz-

funktionen gleich 1 sind, brauchen nur die Koeffizienten von σ_{g1}^2 verglichen zu werden und wir werden nur die Zufallsvariablen $q_1(K)$, $q_2(K, \alpha)$, $q_3(K, \alpha)$ geeignet abschätzen.

3.7.3 Vergleich von θ_{g1}^* und $\theta_{\bar{y}2}^*$

Für $\alpha \geq \frac{1}{(n-1)^2}$, $n \geq 4$ und $2 \leq k \leq n$ gilt:

$$\begin{aligned}
 (3.7.11) \quad q_2(k, \alpha) &\geq \frac{n^2 (n-k) [n-k-1+2(k-2)(n-1)]}{2k(k-1)(n-1)(n-2)(n-3)} \\
 &= q_1(k) \frac{[2(n-k-1)+4(k-2)(n-1)](n-1)}{(K-1)(n-2)(n-3)} \\
 &\geq q_1(k) \left[\frac{2(n-k-1)}{(k-1)(n-3)} + \frac{4(k-2)}{k-1} \right] \\
 &\geq q_1(k) \cdot 2 \geq q_1(k)
 \end{aligned}$$

Da aus Grundgesamtheiten mit weniger als 4 Elementen normalerweise keine Stichprobe gezogen wird, da bei SI $\theta_{\bar{y}2}^*$ für $K \leq 2$ nicht definiert ist und weil $\alpha \geq \frac{1}{(n-1)^2}$ stets gilt, ist also θ_{g1}^* immer besser als $\theta_{\bar{y}2}^*$. Wenn man die Grade beobachten kann, besteht kein Grund, den Stichprobengraphen SI zu wählen. Gründe für die Wahl des Stichprobengraphen SI müssen also bei der Schätzung von $\theta_{\bar{y}2}^*$ darin liegen, daß die Grade der Knoten nicht oder nur sehr fehlerhaft beobachtet werden können, was allerdings bei soziometrischen Anwendungen oft der Fall ist (vgl. 3.6.1 und 3.8).

3.7.4 Vergleich von θ_{g1}^* und $\theta_{\bar{y}3}^*$

Obwohl bei $\theta_{\bar{y}3}^*$ die Information aus dem Stichprobengraphen mehr als bei θ_{g1}^* genutzt wird, ist $\theta_{\bar{y}3}^*$ nicht immer besser als θ_{g1}^* , z.B. bei dem schon erwähnten Fall $\sigma_{g1}^2 = 0$ und $\sigma_{\bar{y}2}^2 > 0$. Wir können allerdings durch Abschätzung von $q_3(K, \alpha)$ eine Grenze angeben, wann $\theta_{\bar{y}3}^*$ besser ist. Für $\alpha \leq \frac{2}{n-1}$, d.h. für $\frac{n-1}{2} \sigma_{\bar{y}2}^2 \leq \sigma_{g1}^2 \leq (n-1)^2 \sigma_{\bar{y}2}^2$ gilt.

$$\begin{aligned}
 (3.7.12) \quad q_3(k, \alpha) &\leq \frac{n^2(n-1)(n-k)(n-k-1)}{2k(2n-k-1)^2(n-2)(n-3)} (2(k-1) + 2(n-k-2)) \\
 &= q_1(k) \frac{4 \cdot (n-1)^2(n-k-1)}{(2n-k-1)^2(n-2)} \\
 &\leq q_1(k)
 \end{aligned}$$

Die letzte Abschätzung erhält man aus der Tatsache, daß $\frac{n-k-1}{(2n-k-1)^2}$ für wachsende k monoton fällt und man daher $k=1$ setzen kann. Wir haben nun das Ergebnis, daß $\theta_{\bar{y}3}^*$ für $\sigma_{g1}^2 \geq \frac{n-1}{2} \sigma_{\bar{y}2}^2$ besser als θ_{g1}^* ist, aber wir haben außer dem Extremfall $\sigma_{g1}^2 = 0$ und $\sigma_{\bar{y}2}^2 > 0$ keine Aussage, wann θ_{g1}^* besser als $\theta_{\bar{y}3}^*$ ist. Um dazu ein Ergebnis zu bekommen, müßte man die Erwartungswerte von $q_1(K)$ und $q_3(K, \alpha)$ betrachten, was bei Termen der Form $E(\frac{1}{f(K)})$ sehr schwierig ist. Da wir bei der obigen Abschätzung als ungünstigsten Fall $k = 1$, d.h. Ziehen eines Knotens, zugrundegelegt haben, kann man sagen daß bei wachsender Stichprobengröße eine immer kleinere untere Grenze für σ_{g1}^2 gewählt werden kann, so daß $\theta_{\bar{y}3}^*$ besser als θ_{g1}^* ist und speziell für $k = n-1$ ist $\theta_{\bar{y}3}^*$ stets mindestens so gut wie θ_{g1}^* .

Um eine bessere Vorstellung von der Bedingung $\sigma_{g1}^2 \geq \frac{n-1}{2} \sigma_{\bar{y}2}^2$ zu bekommen, betrachten wir für feste Werte in der Grundgesamtheit die Kovarianz von $\bar{y}_{ij}$ und $g'_{ij} = g_i - \bar{y}_{ij}$

(man beachte $\sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n g'_{ij} = 2(n-2)t_{\bar{y}2}$):

$$(3.7.13) \quad \sigma_{yg'} = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n (\bar{y}_{ij} - \frac{2}{n(n-1)} t_{\bar{y}2})$$

$$(g'_{ij} - \frac{2(n-2)}{n(n-1)} t_{\bar{y}2}) = \frac{1}{n-1} \sigma_{g1}^2 - \sigma_{\bar{y}2}^2 .$$

Für $\sigma_{yg'}$ lautet die obige Bedingung:

$$(3.7.14) \quad \sigma_{yg'} \geq \frac{1}{n-1} \frac{n-1}{2} \sigma_{\bar{y}2}^2 - \sigma_{\bar{y}2}^2 = -\frac{1}{2} \sigma_{\bar{y}2}^2 ,$$

d.h. wenn die Zahlen $\bar{y}_{ij}$ und g'_{ij} in der Grundgesamtheit nicht zu stark negativ korrelieren, ist $\theta_{\bar{y}3}^*$ besser als θ_{g1}^* .

3.7.5 Bemerkung zum Vergleich von $\theta_{\bar{y}2}^*$ und $\theta_{\bar{y}3}^*$.

Bisher haben wir gezeigt, daß für $\sigma_{g1}^2 \geq \frac{n-1}{2} \sigma_{\bar{y}2}^2$ gilt:

$$(3.7.15) \quad \text{var}(\theta_{\bar{y}2}^*) \geq \text{var}(\theta_{g1}^*) \geq \text{var}(\theta_{\bar{y}3}^*) ,$$

wobei die erste Ungleichung auch für $\sigma_{g1}^2 < \frac{n-1}{2} \sigma_{\bar{y}2}^2$ gilt.

Man kann nun auch zeigen, daß

$$(3.7.16) \quad \text{var}(\theta_{\bar{y}2}^*) \geq \text{var}(\theta_{\bar{y}3}^*)$$

stets gilt, was wir aber nicht beweisen wollen, weil es relativ unwichtig ist. Wenn man für die Knoten der Stichprobe die Grade beobachten kann, wissen wir aus den bisherigen Ergebnissen, daß man sie in der Schätzfunktion auch verwenden und statt $\theta_{\bar{y}2}^*$ entweder θ_{g1}^* oder $\theta_{\bar{y}3}^*$ wählen sollte.

3.7.6 Bemerkung zum Vergleich von SV und SI

Auch wenn Granovetter (1976) sich bei seinem Aufsatz auf die Betrachtung ungerichteter Graphen beschränkt, liegen normalerweise bei soziometrischen Fragestellungen gerichtete Graphen vor. Bei einer Umfrage hat man die Alternativen, jeden Befragten nach der Anzahl seiner Bekannten zu fragen oder ihn seine Bekannten innerhalb der Stichprobe nennen zu lassen. Wenn man die Antworten als Ausgangsgrade und Kantennwerte desselben Graphen auffassen kann (vgl. 3.6.1), würden im ersten Fall die Ausgangsgrade beobachtet und es liegt der Stichprobengraph SV vor, im zweiten Fall liegt der Stichprobengraph SI vor. Ein Vergleich dieser beiden Stichproben erfordert zusätzlich Aussagen über das Verhältnis der Streuung der Ausgangsgrade σ_{a1}^2 und der Grade σ_{g1}^2 . An einem Beispiel wollen wir zeigen, daß die Schätzung θ_{a1}^* aus SV nicht immer besser als $\theta_{\bar{y}2}^*$ aus SI ist. Als Extremfall betrachten wir einen Graphen, bei dem die Streuung σ_{g1}^2 von $G_i = A_i + E_i$ Null beträgt aber $\sigma_{a1}^2 > 0$ gilt, z.B. den in der folgenden Abbildung dargestellten Graphen mit 6 Knoten:

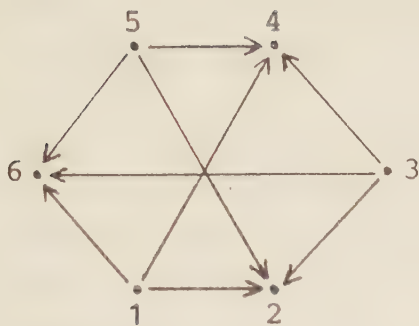


Abb. 3.7.1

$$t_{y2} = t_{\bar{y}2} = 9 \quad g_i = 3, i=1, \dots, 6$$

$$a_i = 3 \quad \text{für } i=1, 3, 5$$

$$a_i = 0 \quad \text{für } i=2, 4, 6$$

$$\sigma_{\bar{y}2}^2 = \frac{2}{30} \cdot 9 - \frac{4}{900} \cdot 9^2 = \frac{24}{100}$$

(Gl. (3.5.14))

$$\sigma_{g1}^2 = 0, \quad \sigma_{a1}^2 = \frac{1}{6} \cdot 6 \cdot \left(\frac{3}{2}\right)^2 = \frac{9}{4}$$

Bei einer symmetrischen Stichprobe von 3 Knoten ($K \equiv k=3$)

$$\text{gilt: } \text{var}(\theta_{a1}^*) = \sigma_{a1}^2 \frac{n^2(n-k)}{k(n-1)} = \frac{9}{4} \cdot \frac{6^2 \cdot 3}{3 \cdot 5} = \frac{81}{5}$$

$$\text{var}(\theta_{\bar{y}2}^*) = \sigma_{\bar{y}2}^2 \frac{n^2(n-1)^2(n-k)(n-k-1)}{k(k-1)^2(n-2)(n-3)}$$

$$= \frac{24}{100} \frac{30^2 \cdot 3 \cdot 2}{3 \cdot 2 \cdot 2 \cdot 4 \cdot 3} = 9 < \frac{81}{5} = \text{var}(\theta_{a1}^*),$$

$\theta_{\bar{y}2}^*$ ist also besser als θ_{a1}^* .

Trotz dieses konstruierten Beispiels wird man allerdings vermuten, daß θ_{a1}^* in den meisten Fällen besser als $\theta_{\bar{y}2}^*$ ist und für $\sigma_{a1}^2 \leq \frac{1}{2}\sigma_{g1}^2$ folgt diese Aussage sofort aus Gl. (3.7.11). Ein genauerer Vergleich überschreitet den Rahmen dieser Arbeit, da dann einschränkendere Modellvorstellungen für die Beziehungen zwischen σ_{a1}^2 , σ_{g1}^2 und $\sigma_{\bar{y}2}^2$ genutzt werden müßten.

3.7.7 Franks Ergebnisse zum Vergleich von Stichprobengraphen

In seinem Buch (1971, S. 108, 138) und dem Übersichtsartikel (1981, S. 140) geht Frank auf Vergleiche von Stichprobengraphen ein. Er beschreibt ein Kostenmodell zur Bewertung von Stichprobengraphen, in dem er Kosten für das Stich-

probenverfahren, die Kosten pro ausgewähltem Knoten und pro beobachteter Kante festlegt. Allerdings führt er dieses Verfahren nicht weiter aus.

Außerdem vergleicht er den Stichprobengraphen SI bei einer Bernoullistichprobe von Knoten mit einer Bernoullistichprobe aus der Menge aller Kanten, wobei die Erwartungswerte $E(K_2)$ der beobachteten Knotenpaare gleich sind. Für die jeweiligen Horvitz-Thompson-Schätzfunktionen und unbewerteten Graphen ist dann die Varianz beim Stichprobengraphen SI kleiner als bei einer Bernoullistichprobe von Kanten. Zu diesem Vergleich ist zu bemerken, daß bei Bernoullistichproben $\text{var}(K)$ bzw. $\text{var}(K_2) > 0$ gilt und die betrachteten Horvitz-Thompson Schätzfunktionen unzulässig sind.

3.8 Bemerkungen zum "network sampling" von Granovetter

3.8.1 Die Motivation des "network sampling"

Wenn man schätzen will, wieviel Leute innerhalb einer Population (z.B. in einem Stadtteil) sich kennen, liegt es zunächst nahe, Personen einer Stichprobe nach der Anzahl ihrer Bekannten im Stadtteil zu fragen. Allerdings muß man vermuten, daß die Personen nur sehr schlecht schätzen werden, auch wenn sie z.B. dazu angehalten werden, zuerst die Namen ihrer Bekannten aufzuschreiben und diese dann zu zählen (vgl. Granovetter (1976)). Wenn man nun den Befragten eine Namensliste vorlegt (bei SI speziell die Namen der Personen der Stichprobe), wird man erwarten können, daß sie beim Ankreuzen ihrer Bekannten zumindest bedeutend weniger Fehler machen. Da es normalerweise für den Befragten unzumutbar ist, eine Liste mit den Namen aller Personen der Population durchzugehen, ermittelt man im zweiten Fall nicht vollständig den Bekanntenkreis der befragten Personen.

Die Fehler beim Nennen von Bekannten wollen wir ignorieren, da sie wegen ihrer Uneinheitlichkeit zu schwer statistisch zu behandeln sind: aus sehr verschiedenen Gründen kann eine Person einen Bekannten nicht nennen wollen oder Personen fälschlich als Bekannte angeben. Wir wollen uns auf das Problem beschränken, daß bei der Frage nach der Anzahl der Bekannten einfach einige vergessen werden.

Falls auf der Liste die Namen der Personen der Stichprobe stehen, führt ein Vergleich beider Fragestellungen zum Vergleich des Stichprobengraphen SI mit dem Stichprobengraphen SV in gerichtetem Fall, wobei die Grade, bzw. Ausgangsgrade im Gegensatz zu den Werten y_{ij} in SI fehlerhaft beobachtet werden. Falls kein Fehler bei der Beobachtung der Grade vorhanden wäre, hätten wir nach den Ergebnissen

des vorigen Abschnittes bei ungerichteten Graphen keine und bei gerichteten Graphen wenig Veranlassung den Stichprobengraphen SI zu wählen. Außer dieser statistischen Begründung sprechen gegen den Stichprobengraphen SI der erhöhte Aufwand für Befragten, Interviewer, Datenauswerter und der größere Eingriff in den persönlichen Bereich des Befragten, der sich vermutlich auch in mehr Antwortverweigerungen niederschlagen wird.

Es würde in diesem Rahmen zu weit führen, die Probleme beim Vergleich beider Fragestellungen vollständig auszuführen; mit den bisherigen Ergebnissen kann man aber Wege zur Behandlung des Problems aufzeigen.

3.8.2 Zur Schätzung des Fehlers bei der Frage nach der Anzahl der Bekannten

Granovetter (1976) stellt fest, daß es keine direkte Evidenz gibt, daß man mit der Frage nach der Anzahl der Bekannten schlechte Ergebnisse erzielt. Diese direkte Evidenz können wir auch nicht geben, aber immerhin ist es möglich zu schätzen, wie sehr sich die Ergebnisse unterscheiden, wenn man einmal summarisch nach der Anzahl der Bekannten fragt und andererseits direkt fragt, welche Personen der Stichprobe Bekannte sind.

Zunächst nehmen wir an, die y_{ij} seien die Elemente der Kantenindikatormatrix eines nicht zufälligen, gerichteten, unbewerteten Graphen. Die y_{ij} zeigen im Beispiel, ob Person i die Person j als Bekannten bezeichnet. Den Knoten ordnen wir Zufallsvariablen

$$(3.8.1) \quad Y_{ii} = a_i - \epsilon_i$$

zu, wobei a_i der Ausgangsgrad des Graphen und ϵ_i eine Zufalls-

variable ist. Im Beispiel sind die Realisierungen von ϵ_i die Zahl der Bekannten, die Person i bei der Frage nach der Anzahl seiner Bekannten vergißt. Da für die Fehlervariablen ϵ_i gilt:

$$(3.8.2) \quad E\left(\frac{1}{n} \sum_{i=1}^n \epsilon_i\right) = \frac{1}{n} \theta_{y1} - \frac{1}{n} \theta_{y2} = \frac{1}{n} \theta_{y1} - \frac{1}{n} \theta_{\bar{y}2}$$

$$(\text{mit } \bar{y}_{ij} = y_{ij} + y_{ji})$$

kann man den durchschnittlichen Fehler ϵ_i pro Person durch die Differenz der Schätzfunktionen $\frac{1}{n} \hat{\theta}_{y1}$ der Knotensumme und $\frac{1}{n} \hat{\theta}_{\bar{y}2}$ der Kantensumme aus den Stichprobengraphen SI schätzen. Schätzungen dieser Art haben wir in Abschnitt 2.5 und 3.5.5 besprochen. Bei der Wahl von $\frac{1}{n} \theta_{y1}^*$ und $\frac{1}{n} \theta_{\bar{y}2}^*$ lautet eine erwartungstreue Schätzfunktion $\hat{\epsilon}$ für

$$E\left(\frac{1}{n} \sum_{i=1}^n \epsilon_i\right):$$

$$(3.8.3) \quad \hat{\epsilon} := \frac{1}{K} \sum_{i=1}^n (y_{ii} - \hat{A}_i) S_i$$

wobei $\hat{A}_i := \frac{n-1}{K-1} \sum_{\substack{j=1 \\ j \neq i}}^n y_{ij} S_j$ eine Schätzfunktion für den Ausgangsgrad a_i des Knotens i aus dem Stichprobengraphen SI ist.

Die Schätzung weiterer Parameter der Verteilungen der ϵ_i ist mit den Ergebnissen dieser Arbeit nicht direkt möglich, allerdings erhält man aus den ersten beiden Kapiteln erwartungstreue Schätzfunktionen σ_{g1}^{2*} (Gl. (2.4.5)) für die Streuung der Grade $g_i = a_i + \epsilon_i$, σ_{y1}^{2*} (Gl. (1.5.5)) für die Streuung der

Y_{ii} und σ_{3yg}^* (Gl. (2.5.22)) für die Abhängigkeit der Y_{ii} und der Grade g_i .

Wenn man ungerichtete Graphen betrachtet und Y_{ii} durch $Y_{ii} = g_i - \varepsilon_i$ definiert, gilt für die Streuung σ_ε^2 der Fehler ε_i :

$$\begin{aligned} (3.8.4) \quad \sigma_\varepsilon^2 &= E\left(\frac{1}{n} \sum_{i=1}^n (\varepsilon_i - \frac{1}{n} \sum_{i=1}^n \varepsilon_i)^2\right) \\ &= \sigma_{y1}^2 + \sigma_{g1}^2 - 2\sigma_{3yg} \end{aligned}$$

und man kann aus den erwartungstreuen Schätzfunktionen für σ_{y1}^2 , σ_{g1}^2 und σ_{3yg} eine erwartungstreue Schätzfunktion für σ_ε^2 bilden.

Bisher sind wir davon ausgegangen, daß man die Fehler ε_i nicht direkt beobachtet. Es ist allerdings durchaus möglich, zu Kanten innerhalb des Stichprobengraphen SI Fehlervariablen ε_{ij} zu beobachten, wenn man z.B. zunächst den Befragten eine Liste seiner Bekannten erstellen läßt und ihn dann auf einer Liste mit den Personen der Stichprobe seine Bekannten ankreuzen läßt. Beim Vergleich beider Listen beobachtet man für Personenpaare der Stichprobe die Abweichungen ε_{ij} direkt und hat bei der Schätzung ihrer Summe oder des Mittels das in 3.5 behandelte Problem der Schätzung der Kantensumme eines Graphen aus der Stichprobe SI.

3.8.3 Zum Vergleich der Schätzfunktionen θ_{y1}^* und $\theta_{\bar{y}2}^*$

In dem vorigen Unterabschnitt haben wir das Mittel der Abweichungen ε_i geschätzt; jetzt wollen wir umgekehrt

die Auswirkungen der Fehlervariablen auf die Varianz von θ_{Y1}^* im Vergleich zu $\theta_{\bar{Y}2}^*$ untersuchen und speziell auf die Frage eingehen, ob bzw. wann sich das schwerer durchzuführende Beobachtungsverfahren SI überhaupt lohnt. Dabei wollen wir wie Granovetter unbewertete, ungerichtete Graphen betrachten und die Problematik bei gerichteten Graphen nur kurz ansprechen.

Ausgangspunkt für ein Fehlermodell ist wieder ein nicht zufälliger, ungerichteter, unbewerteter Graph mit Kantenindikatorvariablen $\bar{y}_{ij}$. Sie zeigen im Beispiel an, ob Person i und j Bekannte sind und wir nehmen an, daß sie im Stichprobengraphen SI fehlerfrei beobachtet werden.

Den Knoten ordnen wir die Zufallsvariablen

$$(3.8.5) \quad Y_{ii} = g_i - \epsilon_i = g_i - \sum_{\substack{j=1 \\ j \neq i}}^n \epsilon_{ij}$$

zu, wobei g_i der Grad des Knotens i ist. Im Beispiel sind die Realisierungen dieser Zufallsvariablen die Antwort auf die Frage nach der Anzahl der Bekannten und die Zufallsvariable ϵ_{ij} zeigt an, ob der Bekannte j von Person i vergessen wurde. Weil wir uns nur auf das Problem beschränkt hatten, daß Bekannte vergessen werden, gilt $\epsilon_{ij} \equiv 0$ für $\bar{y}_{ij} = 0$. In Meßfehlermodellen ist es plausibel anzunehmen, daß die Fehler ϵ_{ij} stochastisch unabhängig sind (vgl. Abschnitt 3.2). Außerdem setzen wir die ϵ_{ij} für $\bar{y}_{ij} = 1$ als identisch verteilte Bernoullivariablen voraus und definieren:

$$(3.8.6) \quad \alpha := P(\epsilon_{ij}=1) = 1 - P(\epsilon_{ij}=0)$$

Die Beziehungen zwischen Parametern der Verteilung der Y_{ii} und des zugrundeliegenden festen Graphen werden durch folgende Gleichungen beschrieben:

$$(3.8.7) \quad E(Y_{ii}) = g_i(1-\alpha)$$

$$(3.8.8) \quad E(Y_{ii}^2) = E^2(Y_{ii}) + \text{var}\left(\sum_{\substack{j=1 \\ j \neq i}}^n \epsilon_{ij}\right) \\ = g_i^2(1-\alpha)^2 + g_i\alpha(1-\alpha)$$

$$(3.8.9) \quad \theta_{Y1} = E\left(\sum_{i=1}^n Y_{ii}\right) = 2\theta_{\bar{Y}2}(1-\alpha)$$

$$(3.8.10) \quad \text{var}(T_{Y1}) = \text{var}\left(\sum_{i=1}^n \sum_{j=i}^n \epsilon_{ij}\right) \\ = 2\theta_{\bar{Y}2}\alpha(1-\alpha)$$

$$(3.8.11) \quad \sigma_{Y1}^2 = \frac{1}{n} \sum_{i=1}^n E(Y_{ii}^2) - \frac{1}{n^2} (\text{var}(T_{Y1}) + \theta_{Y1}^2) \\ = \frac{1}{n} (1-\alpha)^2 \sum_{i=1}^n g_i^2 + \frac{2}{n} \theta_{\bar{Y}2} \alpha(1-\alpha) \\ - \frac{2}{n^2} \theta_{\bar{Y}2}^2 \alpha(1-\alpha) - \frac{4}{n^2} \theta_{\bar{Y}2}^2 (1-\alpha)^2 \\ = \sigma_{g1} (1-\alpha)^2 + \frac{2(n-1)}{n^2} \theta_{\bar{Y}2}^2 \alpha(1-\alpha)$$

(vgl. Gl. (1.3.15) und (2.2.21) mit Berücksichtigung, daß $\text{var}(\bar{y}_{ij})=0$).

Wir wollen nun die Anzahl der Bekannten pro Person, d.h. $\frac{2}{n}\theta_{\bar{y}2}$ schätzen und vergleichen die aus den Y_{ii} gebildete Schätzfunktion $\frac{1}{n}\theta_{y1}^*$ mit $\frac{2}{n}\theta_{\bar{y}2}^*$. Für die quadrierten Abweichungen vom wahren Wert gilt nach Satz 1.3.5 und 2.2.5:

$$\begin{aligned}
 (3.8.12) \quad E\left(\frac{1}{n}\theta_{y1}^* - \frac{2}{n}\theta_{\bar{y}2}\right)^2 &= \frac{1}{n^2}E(\theta_{y1}^* - \theta_{y1})^2 \\
 &\quad + \left(\frac{1}{n}\theta_{y1} - \frac{2}{n}\theta_{\bar{y}2}\right)^2 \\
 &= \sigma_{y1}^2 E\left(\frac{n-K}{K(n-1)}\right) + \frac{1}{n^2}\text{var}(T_{y1}) \\
 &\quad + \frac{4}{n^2}\theta_{\bar{y}2}^2 \alpha^2 \\
 &= \sigma_{g1}^2 (1-\alpha)^2 E\left(\frac{n-K}{K(n-1)}\right) \\
 &\quad + \frac{2}{n}\theta_{\bar{y}2} \alpha (1-\alpha) E\left(\frac{(n-K)}{nK}\right) \\
 &\quad + \frac{2}{n}\theta_{\bar{y}2} \alpha (1-\alpha) + \frac{4}{n^2}\theta_{\bar{y}2}^2 \alpha^2 \\
 &= \sigma_{g1}^2 (1-\alpha)^2 E\left(\frac{n-K}{K(n-1)}\right) \\
 &\quad + \frac{2}{n}\theta_{\bar{y}2} \alpha (1-\alpha) E\left(\frac{1}{K}\right) \\
 &\quad + \frac{4}{n^2}\theta_{\bar{y}2}^2 \alpha^2
 \end{aligned}$$

$$(3.8.13) \quad E\left(\frac{2}{n}\theta_{\bar{y}2}^* - \frac{2}{n}\theta_{\bar{y}2}\right)^2 = \sigma_{\bar{y}2}^2 E\left(\frac{2(n-1)^2(n-K)(n-K-1)}{(n-2)(n-3)K(K-1)}\right)$$

$$\begin{aligned}
 & + \sigma_{g1}^2 E \left(\frac{4(n-1)(K-2)(n-K)}{(n-2)(n-3)K(K-1)} \right) \\
 & = \sigma_{g1}^2 E \left(\frac{4(n-1)(K-2)(n-K)}{(n-2)(n-3)K(K-1)} \right) \\
 & + \frac{2}{n} \theta_{\bar{y}2} E \left(\frac{2(n-1)(n-K)(n-K-1)}{(n-2)(n-3)K(K-1)} \right) \\
 & - \frac{4}{n} \theta_{\bar{y}2}^2 E \left(\frac{2(n-K)(n-K-1)}{(n-2)(n-3)K(K-1)} \right)
 \end{aligned}$$

(vgl. Gl. (3.5.7) mit $\text{var}(T_{\bar{y}2})=0$)

Für $\alpha=0$ (d.h. Y_{ii} sind die tatsächlichen Grade) haben wir in 3.7.3 gezeigt, daß θ_{y1}^* stets besser als $\theta_{\bar{y}2}^*$ ist; falls $\alpha>0$ ist, hängt es von dem Verhältnis von σ_{g1}^2 und $\theta_{\bar{y}2}$ ab, welche Schätzfunktion besser ist. Da σ_{g1}^2 bei θ_{y1}^* bedeutend geringer gewichtet wird als bei $\theta_{\bar{y}2}^*$, ist also bei großen σ_{g1}^2 zu erwarten, daß θ_{y1}^* die kleinere quadrierte Abweichung hat, während bei großen $\theta_{\bar{y}2}$ eher der Stichprobengraph SI zu bevorzugen ist.

Die folgenden Kurven geben exemplarisch eine Vorstellung, wie zu vorgegebenen Werten von α das Verhältnis von Standardabweichung σ_{g1} und dem Mittel $\frac{2}{n} \theta_{\bar{y}2}$ der Grade sein muß, um mit dem Stichprobengraphen SI eine bessere Schätzfunktion als bei der fehlerhaften Beobachtung der Grade zu erhalten. Bei der ersten Abbildung sind wir von einer Grundgesamtheit von $n = 1000$ Untersuchungseinheiten, bei der zweiten Abbildung von $n = 10000$ Untersuchungseinheiten ausgegangen und haben eine Stichprobe vom Umfang $k = 100$ gezogen. Für verschiedene Werte von α ordnen die Kurven den Mitteln $\frac{2}{n} \theta_{\bar{y}2}$ der Grade den Wert der Streuung σ_{g1}^2 zu, bei dem der Erwartungswert der quadrierten Abweichung vom wahren Wert für beide Schätzverfahren gleich ist.

Für Parameter $(\sigma_{g1}, \frac{2}{n}\bar{\theta}_{\bar{y}2})$ oberhalb der Kurve ist θ_{g1}^* besser als $\theta_{\bar{y}2}^*$, unterhalb der Kurve ist $\theta_{\bar{y}2}^*$ besser. Aus Abb. 3.8.1 können wir z.B. für $\alpha \leq 0,1$ und $\frac{2}{n}\bar{\theta}_{\bar{y}2} < 15$ schließen, daß die Beobachtungsregel SI zu keiner besseren Schätzung führt. Die obigen Voraussetzungen können bei soziometrischen Fragen nach relativ engen Freundschaftsbeziehungen durchaus realistisch sein. Für $\alpha=0,2$ sieht man, daß der Variationskoeffizient im angegebenen Bereich etwas größer als 1 sein muß, damit θ_{g1}^* besser ist. Im Fall der bedeutend größeren Grundgesamtheit (Abb. 3.8.2) lohnt sich das Beobachtungsverfahren SI schon nicht mehr, wenn man z.B. von einem durchschnittlichen Bekanntenkreis kleiner als 45 Personen ausgeht und man annimmt, daß höchstens 20 % bei der Frage nach der Anzahl vergessen werden.

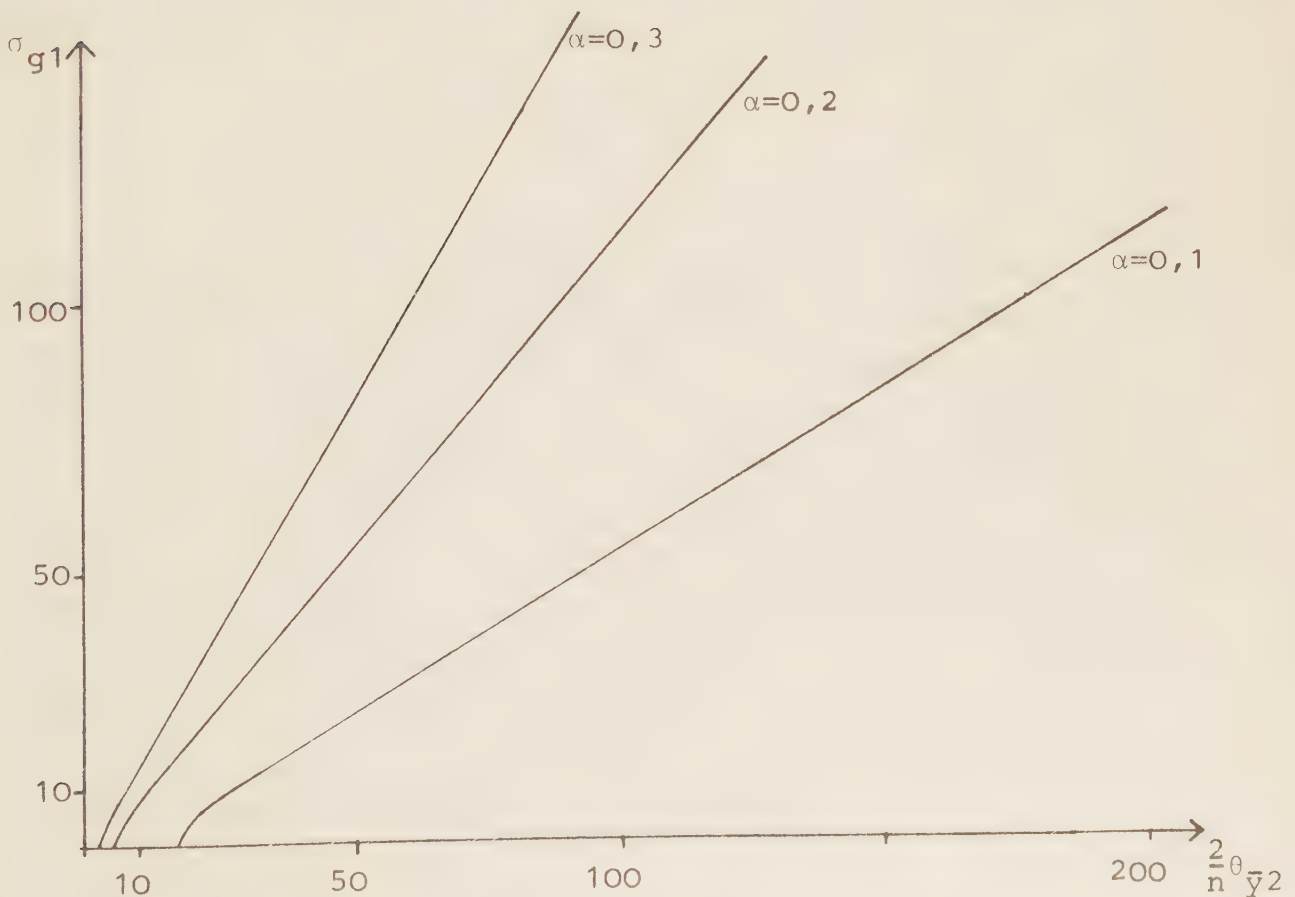


Abb. 3.8.1

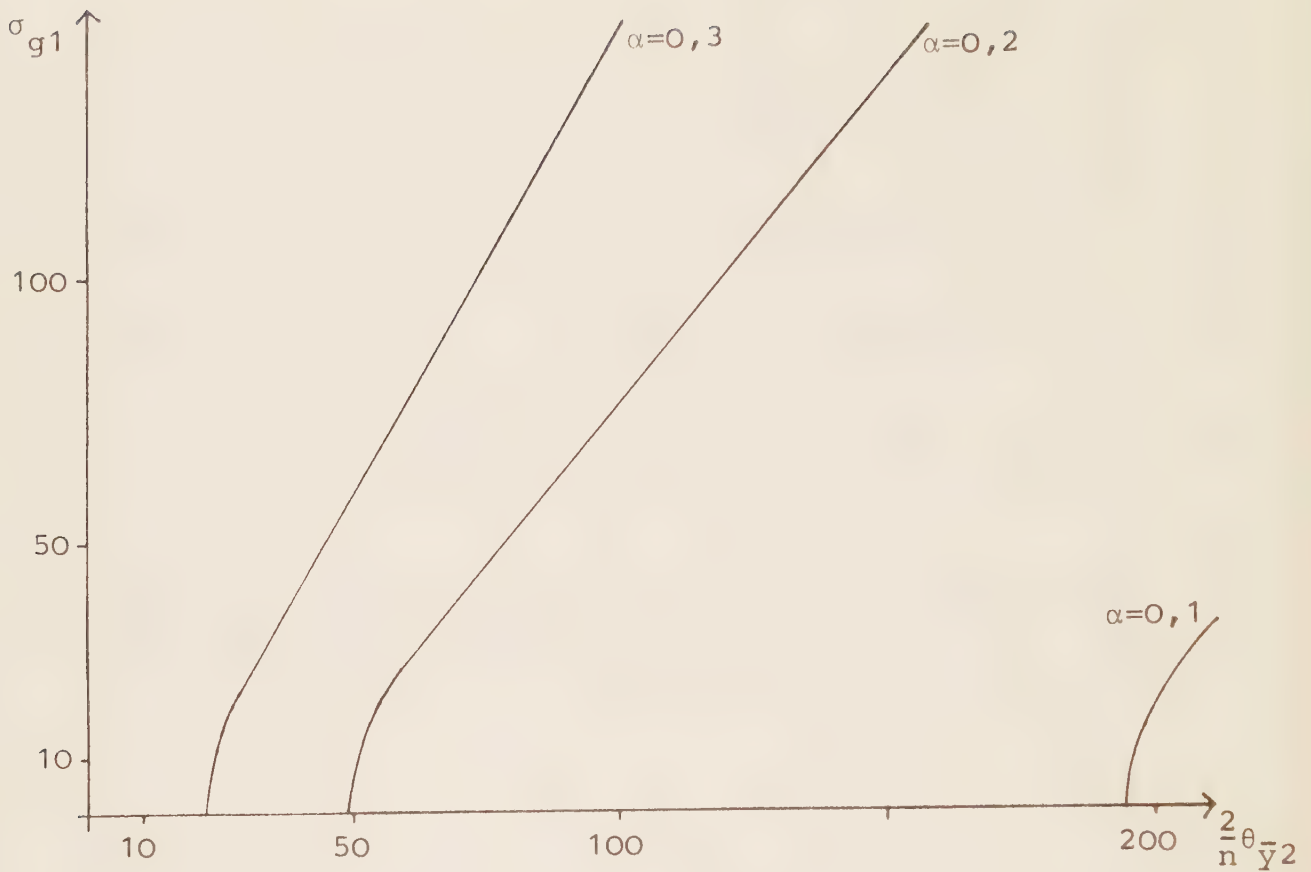


Abb. 3.8.2

Abschließend wollen wir auf Probleme bei der Behandlung von gerichteten Graphen hinweisen, die bei soziometrischen Fragestellungen oft realistischer sind. In diesem Fall wird man anstelle der Grade die Ausgangsgrade der Knoten fehlerhaft beobachten. Man erhält $E(\frac{1}{n}\theta_{y1}^* - \frac{2}{n}\theta_{\bar{y}2})^2$ als Funktion von $\sigma_a^2, \theta_{\bar{y}2}$ und α und $E(\frac{2}{n}\theta_{\bar{y}2}^* - \theta_{\bar{y}2})^2$ als Funktion von $\sigma_g^2, \theta_{\bar{y}2}$ und der Anzahl der Doppelkanten (vgl. Gl. (3.5.14)). Bei einem Vergleich hat man also das Verhältnis von 5 Parametern zu berücksichtigen anstelle von nur 3 Parametern im ungerichteten Fall.

3.8.4 Weitere Bemerkungen zum "network sampling"

In Abschnitt 3.3.7 haben wir schon gesagt, daß Granovetter (1976) und Erickson, Nosanchuk, Lee (1981) nicht von einem Stichprobengraphen SI ausgehen, sondern zumindest in ihren Beispielen andere Stichprobengraphen wählen.

Da Erickson u.a. weniger an Inferenzproblemen, sondern an Problemen der praktischen Durchführung des "network sampling" interessiert sind, beschreiben sie die Art ihres Stichprobengraphen nicht sehr exakt. Aus ihren Ausführungen geht allerdings hervor, daß sie bei einem Vortest getrennte Stichproben für die Befragten $S^{(B)}$ und für die Namensliste $S^{(L)}$ wählen. Wenn beide Stichproben unabhängig voneinander mit einem symmetrischen Stichprobenverfahren gezogen werden, können wir den resultierenden Stichprobengraphen auch mit den bisherigen Ergebnissen behandeln. Wenn die Namensliste durch Auswahl von $k^{(L)}$ Namen ohne Zurücklegen entsteht, ist für jeden Knoten i :

$$\hat{A}_i = \frac{n-1}{k^{(L)} - S_i^{(L)}} \sum_{\substack{j=1 \\ j \neq i}}^n y_{ij} S_j^{(L)}$$

eine erwartungstreue Schätzfunktion für den Ausgangsgrad A_i , die vom Stichprobenverfahren zur Wahl von $S^{(B)}$ unabhängig ist und die Schätzung von θ_{y2} können wir als Schätzung des Erwartungswertes der Summe der $\hat{A}_i$ auffassen. Somit haben wir das Problem von Kapitel I bzw. des Stichprobengraphen SV mit Beobachtungen von Realisierungen von Zufallsvariablen $\hat{A}_i$. Es muß angemerkt werden, daß die a_{11} , σ_1^2 usw. entsprechenden Parameter nun nicht Parameter des zugrundeliegenden Graphen, sondern der Verteilung der Schätzfunktionen $(\hat{A}_1, \dots, \hat{A}_n)$ sind.

Um die Namensliste zumutbar klein zu halten, wählt Granovetter (1976, S. 1298) ein für Berechnungen und für die

praktische Durchführung schwieriges Verfahren, das mit den Mitteln dieser Arbeit nicht direkt zu behandeln ist. Wenn er z.B. eine Stichprobe von 500 Befragten auswählt, nummeriert er sie durch und befragt den ersten zu den Personen 2 - 251, den zweiten zu Nr. 3 - 252, ..., den Befragten Nr. 500 zu Nr. 1 - 249. Das Stichprobenverfahren für die Liste ist also abhängig von der Auswahl der Befragten ($S^{(L)} cS^{(B)}$). Außerdem sind die (zufällig aufgestellten) Listen, die Person i und Person j der Stichprobe vorgelegt werden, auch abhängig voneinander und es ist recht kompliziert, die Verteilung der Stichprobenindikatorvariablen S_{ij} für die Kanten zu bestimmen. Da dieses Schema auch nicht sehr praktisch ist (man muß für jeden Befragten eine andere Namensliste erstellen), behandeln wir es nicht weiter.

Abschließend wollen wir auf zwei Probleme des "network sampling" hinweisen, die von Granovetter nicht angesprochen werden und in dieser Arbeit auch nur am Rande erwähnt werden. Empirisch findet man in der Soziometrie normalerweise gerichtete (oder signierte) Graphen vor (z.B. bei der Frage nach Freunden oder Bekannten), auch wenn das, was gemessen wird, eine symmetrische Beziehung ist. Der für die Motivation des Stichprobengraphen SI wichtige Vergleich der Frage nach der Anzahl von Bekannten mit der Nennung von Bekannten gestaltet sich beim realistischeren Modell des gerichteten Graphen wesentlich komplizierter als im gerichteten Fall, der von Granovetter ausschließlich und in dieser Arbeit überwiegend behandelt wird (vgl. 3.7.6 und 3.8).

Wenn man symmetrische Beziehungen untersucht (z.B. haben Person i und j sich besucht) und Personen danach fragt, ist die Darstellung des empirischen Befundes als unbewerteter, ungerichteter Graph oft problematisch. Im Stichprobengraphen SI hat man zur Fragestellung, ob i und j sich besucht haben, zwei möglicherweise widersprüchliche Antworten: die der

Person i und die der Person j . Zumindest bei widersprüchlichen Angaben steht man vor dem Problem, welchen Wert man dem beobachteten Paar (i,j) zuordnet und was für Auswirkungen bestimmte Zuordnungsregeln auf die Schätzungen haben. Das in 3.2.3 und 3.8.3 genannte Meßfehlermodell bietet eine Möglichkeit zur Behandlung widersprüchlicher Daten, man beobachtet 2 Realisierungen der Zufallsvariablen $Y_{ij} = \mu_{ij} + \varepsilon_{ij}$, wobei μ_{ij} der wahre Wert und ε_{ij} der Meßfehler ist.

LITERATURVERZEICHNIS

- BERGE, C. (1962) - The theory of graphs
Methuen, London
- BERNARD, H.R., KILLWORTH, P.D., SAILER, L. (1980) - Informant
accuracy in social network data IV: a comparison of
clique-level structure in behavioral and cognitive
network data
Social Networks, 2, S. 191 - 218
- BERNARD, H.R., KILLWORTH, P.D., SAILER, L. (1981) - A note
on inferences regarding network subgroups - response
to Burt and Bittner
Social Networks, 3, S. 89 - 92
- BIESS, G. (1976) - Graphentheorie
Teubner, Leipzig
- BIYANI, S.M. (1980) - On some estimation problems in finite
population sampling
University Microfilms International, Ann Arbor
- BLACKWELL, D., GIRSHICK, M.A. (1954) - Theory of games and
statistical decisions
Wiley, New York
- BURT, R.S. (1978) - Applied network analysis
Sociological Methods & Research, Vol. 7, No. 2,
S. 123 - 129
- BURT, R.S. (1980) - Models of network structure
Annual Review of Sociology, Vol. 6, S. 79 - 141

- BURT, R.S., BITTNER, Wm.M. (1981) - A note on inferences
regarding network subgroups
Social Networks, 3, S. 71 - 88
- CAPOBIANCO, M. (1972) - Estimating the connectivity of a
graph. In Y. Alavi, D.R. Lick, A.T. White (Eds) -
Graph theory and applications
Springer, Berlin
- COCHRAN, W.G. (1963) - Sampling techniques
Wiley, New York
- ERICKSON, B.H., NOSANCHUK, T.A., LEE, E. (1981) - Network
sampling in practice: some second steps
Social Networks, 3, S. 127 - 136
- FELLER, W. (1962) - An introduction to probability theory
and its applications. Vol. I
Wiley, New York
- FELLER, W. (1971) - An introduction to probability theory
and its applications. Vol. II
Wiley, New York
- FELLING, A.J.A. (1975) - A graph-theoretical approach to
the structure of local elites
Zeitschrift für Soziologie, Jg. 4, Heft 3, S. 221
- 233
- FERGUSON, T.S. (1967) - Mathematical statistics
Academic Press, New York
- FISCHER, C.S. (1982) - What do we mean by 'friend'? An
inductive study.
Social Networks, 3, S. 287 - 306

- FLAMENT, C. (1963) - Applications of graph theory to group structures
Prentice-Hall, Englewood Cliffs
- FRANK, O. (1969) - Structure inference and stochastic graphs
FOA-Reports, 3, S. 1 - 8
- FRANK, O. (1971) - Statistical inference in graphs
Forsvarets Forskningsanstalt,
Stockholm
- FRANK, O. (1977a) - Estimation of graph totals
Scandinavian Journal of Statistics, 4, S. 81 - 89
- FRANK, O. (1977b) - A note on Bernoulli sampling in graphs
and Horvitz-Thompson estimation
Scandinavian Journal of Statistics, 4, S. 178 - 180
- FRANK, O. (1978a) - Sampling and estimation in large social
networks
Social Networks, 1, S. 91 - 101
- FRANK, O. (1978b) - Estimation of the number of connected
components in a graph by using a sampled subgraph
Scandinavian Journal of Statistics, 5, S. 177 - 188
- FRANK, O. (1979a) - Estimating a graph from triad counts
Journal of Statistical Computation and Simulation,
9, S. 31 - 46
- FRANK, O. (1979b) - Moment properties of subgraph counts in
stochastic graphs
Annals of the New York Academy of Sciences, 319,
S. 207 - 218

- FRANK, O. (1980) - Sampling and inference in a population graph
International Statistical Review, 48, S. 33 - 41
- FRANK, O. (1981) - A survey of statistical methods for graph analysis. In S. Leinhardt (Ed)-Sociological methodology 1981
Jossey-Bass, San Francisco
- FREEMAN, L.C. (1976) - Bibliography on social networks
Council of Planning Librarians, Montecello
- FRIEDKIN, N.E. (1981) - The development of structure in random networks: an analysis of the effects of increasing network density on five measures of structure
Social Networks, 3, S. 41 - 52
- GOODMAN, L.A. (1961) - Snowball sampling
Annals of Mathematical Statistics, 32, S. 148 - 170
- GRANOVETTER, M. (1976) - Network sampling: some first steps.
American Journal of Sociology, 81, S. 1287 - 1303
- GRANOVETTER, M. (1977) - Reply to Morgan and Rytina
American Journal of Sociology, 83, S. 727 - 729
- HARARY, F. (1969) - Graph theory
Addison-Wesley, Reading
- HARARY, F., NORMAN, R.Z., CARTWRIGHT, D. (1965) - Structural models: an introduction to the theory of directed graphs
Wiley, New York

HOLLAND, P., LEINHARDT, S. (1975) - Local structure in social networks. In D.R. Heise (Ed.) - Sociological Methodology 1976
Jossey-Bass, San Francisco

HOLLAND, P. LEINHARDT, S. (1981) - An exponential family of probability distributions for directed graphs
Journal of the American Statistical Association, 76, S. 33 - 50

IOFFE, A.D., TICHOMIROV, V.M. (1979) - Theorie der Extremalaufgaben
Deutscher Verlag der Wissenschaften, Berlin

KÄHLER, H.D. (1975) - Das Konzept des sozialen Netzwerkes: Eine Einführung in die Literatur
Zeitschrift für Soziologie, Jg. 4, Heft 3, S. 283 - 290

KLIEMANN, W., MÜLLER, N. (1976) - Logik und Mathematik für Sozialwissenschaftler 2
W. Fink Verlag, München

KRICKEBERG, K. (1963) - Wahrscheinlichkeitstheorie
Teubner, Stuttgart

LUENBERGER, D.G. (1973) - Introduction to linear and non-linear programming
Addison-Wesley, Reading

MCCALLISTER, C.M., FISCHER, C.S. (1978) - A procedure for surveying personal networks
Sociological Methods & Research, 7, S. 131 - 148

- MORENO, J.L. (1974) - Die Grundlagen der Soziometrie
Westdeutscher Verlag, Opladen
- MORGAN, D.L., RYTINA, S. (1977) - Comment on 'network sampling: some first steps' by Mark Granovetter
American Journal of Sociology, 83, S. 722 - 727
- NEHNEVAJSA, J. (1973) - Soziometrie. In König, R. (Hrsg.)
Handbuch der empirischen Sozialforschung, Band 2,
F. Enke Verlag, Stuttgart
- NIEMÖLLER, K., SCHIJF, B. (1980) - Applied network analysis
Quality and Quantity, 14, S. 101 - 116
- ROY, J., CHAKRAVARTI, J.M. (1960) - Estimating the mean of
a finite population
Annals of Mathematical Statistics, 31, S. 392 - 398
- ROYALL, R.M. (1970) - On finite population sampling theory
under certain linear regression models
Biometrika, 57, S. 377 - 387
- SINHA, B.K. (1976) - On balanced sampling schemes
Calcutta Statistical Association Bulletin, 25,
S. 129 - 138
- SNIJERS, T.A.B. (1981) - The degree variance: an index of
graph heterogeneity
Social Networks, 3, S. 163 - 174
- SØRENSEN, A.B., SØRENSEN, A. (1975) - Mathematical sociology
Current Sociology, XXIII, No. 3
- STENGER, H. (1971) - Stichprobentheorie
Physica-Verlag, Würzburg

- STENGER, H. (1972a) - Symmetriebegriffe in der Stichprobentheorie
Beiträge zur angewandten Wirtschaftsforschung,
Mannheim, discussion paper No. 30/1972
- STENGER, H. (1972b) - Effizienzvergleiche zwischen Stichprobenverfahren
Statistische Hefte, 13, S. 314 - 334
- STENGER, H. (1975) - Admissibility and completeness in sampling for proportions or other characteristics with restricted range
Beiträge zur angewandten Wirtschaftsforschung,
Mannheim, discussion paper No. 68/1975
- STENGER, H. (1979) - Loss functions and admissible estimators in survey sampling
Metrika, 26, S. 205 - 214
- STENGER, H., GABLER, S. (1981) - On the completeness of the class of fixed size sampling strategies
Annals of Statistics, 9, No. 1, S. 229 - 232
- STOKMAN, F.N., VAN VEEN, F.J.A.M. (Ed) (1980) - GRADAP user's manual
University of Amsterdam
- WALD, A. (1950) - Statistical decision functions
Wiley, New York
- WASSERMAN, S.S. (1977) - Random directed graph distributions and the triad census in social networks
Journal of Mathematical Sociology, 5, S. 61 - 86
- WITTING, H. (1974) - Mathematische Statistik
Teubner, Stuttgart

ANHANG

Vergleich von Bezeichnungen aus Kapitel I, II und III

In der Arbeit wurde für die theoretischen Kapitel I und II eine etwas andere Bezeichnungsweise als in Kapitel III bei der Anwendung der Ergebnisse auf Graphen gewählt. Erläuterungen der Notation finden sich in der Arbeit in den Abschnitten 2.1 und 3.3. Dieser Anhang soll eine zusätzliche Hilfe zum Vergleich der Bezeichnungen bieten. Nicht berücksichtigt wird hier die Bezeichnungsweise im Fall einer symmetrischen Stichprobe von Knotenpaaren; dieser Vergleich befindet sich auf Seite 163 - 165.

A1. Bezeichnungen von Zufallsvariablen

in Kap. I und II	bei Anwendung auf Graphen (Kap. III)
X_i Zufallsvariablen, die Elementen i der Grundgesamtheit zugeordnet sind	$Y_{ii}(\bar{Y}_{ii})$ Variablen, die Knoten i bei (un)-gerichteten Graphen Werte zuordnen
X_{ij} Variablen, die geordneten Tupeln (i,j) zugeordnet sind	A_i, E_i, G_i Ausgangsgrad, Eingangsgrad und Grad von i
$\bar{X}_{ij}$ Variablen, die ungeordneten Paaren $\{i,j\}$ zugeordnet sind	Y_{ij} Variablen, die gerichteten Kanten Werte zuordnen
	$\bar{Y}_{ij}$ Variablen, die ungerichteten Kanten Werte zuordnen
	$\bar{Y}_{ij} = Y_{ij} + Y_{ji}$ in gerichteten Graphen
$U_i = \sum_{\substack{j=1 \\ j \neq i}}^n \bar{X}_{ij}$	G_i mit $\bar{X}_{ij} = Y_{ij}$ Grad von Knoten i
$T_1 = \sum_{i=1}^n X_i$	$T_{Y1}(T_{\bar{Y}1})$ mit $X_i = Y_{ii}(\bar{Y}_{ii})$
	T_{a1} mit $X_i = A_i$, T_{e1} mit $X_i = E_i$,
	T_{g1} mit $X_i = G_i$
$T_2 = \sum_{i=1}^n \sum_{j>i}^n \bar{Y}_{ij}$	$T_{\bar{Y}2}$ mit $\bar{X}_{ij} = \bar{Y}_{ij}$ in ungerichteten Graphen
	$T_{Y2} = \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n Y_{ij} = T_{\bar{Y}2}$ mit $\bar{Y}_{ij} = Y_{ij} + Y_{ji}$ in gerichteten Graphen
$S_i = \begin{cases} 1 & \text{falls } i \in S \\ 0 & \text{sonst} \end{cases}$	S_i
$S_{ij} = S_i S_j$	$S_{ij} = S_i S_j$ für Stichprobengraphen SI
	$S_{ij} = S_i$ für Stichprobengraphen SV
	$S_{ij} = S_j$ für Stichprobengraphen SN
	$S_{ij} = S_i + S_j - S_i S_j$ für " " SVN
K Stichprobengröße	K Anzahl der beobachteten Knoten

A2. Bezeichnungen von Parametern und ihren Schätzfunktionen

in Kap. I	in Kap. III bei Einsetzen für X_i von				
	Y_{ii}	$\bar{Y}_{ii}$	$E_i^{(1)}$	$A_i^{(2)}$	$G_i^{(3)}$
$\theta_1 = E(T_1) = E\left(\sum_{i=1}^n X_i\right)$	θ_{y1}	$\theta_{\bar{y}1}$	θ_{e1}	θ_{a1}	θ_{g1}
$a_{11} = E\left(\sum_{i=1}^n X_i^2\right)$	a_{y11}	$a_{\bar{y}11}$	a_{e11}	a_{a11}	a_{g11}
$a_{12} = E\left(\sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n X_i X_j\right)$	a_{y12}	$a_{\bar{y}12}$	a_{e12}	a_{a12}	a_{g12}
$\sigma_1^2 = E\left(\frac{1}{n} \sum_{i=1}^n \left(X_i - \frac{1}{n} T_1\right)^2\right)$	σ_{y1}^2	$\sigma_{\bar{y}1}^2$	σ_{e1}^2	σ_{a1}^2	σ_{g1}^2
$\sigma_{11}^2 = \frac{1}{n} \sum_{i=1}^n \left(E(X_i) - \frac{1}{n} \theta_1\right)^2$	σ_{y11}^2	$\sigma_{\bar{y}11}^2$	σ_{e11}^2	σ_{a11}^2	σ_{g11}^2
$\hat{\theta}_1 = \beta(K) \sum_{i=1}^n X_i S_i$	$\hat{\theta}_{y1}$	$\hat{\theta}_{\bar{y}1}$	$\hat{\theta}_{e1}$	$\hat{\theta}_{a1}$	$\hat{\theta}_{g1}$
$\theta_1^* = \frac{n}{K} \sum_{i=1}^n X_i S_i$	θ_{y1}^*	$\theta_{\bar{y}1}^*$	θ_{e1}^*	θ_{a1}^*	θ_{g1}^*
$\tilde{\theta}_1 = \frac{n}{E(K)} \sum_{i=1}^n X_i S_i$	$\tilde{\theta}_{y1}$	$\tilde{\theta}_{\bar{y}1}$	$\tilde{\theta}_{e1}$	$\tilde{\theta}_{a1}$	$\tilde{\theta}_{g1}$
$a_{11}^* = \frac{n}{K} \sum_{i=1}^n X_i^2 S_i$	a_{y11}^*	$a_{\bar{y}11}^*$	a_{e11}^*	a_{a11}^*	a_{g11}^*
$a_{12}^* = \frac{n(n-1)}{K(K-1)} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n X_i X_j S_i S_j$	a_{y12}^*	$a_{\bar{y}12}^*$	a_{e12}^*	a_{a12}^*	a_{g12}^*
$\sigma_1^{2*} = \frac{n-1}{n^2} a_{11}^* - \frac{1}{n^2} a_{12}^*$	σ_{y1}^{2*}	$\sigma_{\bar{y}1}^{2*}$	σ_{e1}^{2*}	σ_{a1}^{2*}	σ_{g1}^{2*}

Anmerkung: Die Schätzfunktionen sind nur für die Stichprobengraphen
⁽¹⁾ SN und SVN, ⁽²⁾ SV und SVN, ⁽³⁾ SVN definiert

in Kap. II	in Kap. III bei Einsetzen von $\bar{y}_{ij}$ für $\bar{x}_{ij}$	Schätzfunktionen für die Parameter (*)
$\theta_2 = E(T_2) = E(\sum_{i=1}^n \sum_{j>1}^n \bar{x}_{ij})$	$\theta_{\bar{y}2} = \theta_{y2}$ mit $\bar{y}_{ij} = y_{ij} + y_{ji}$	$\hat{\theta}_{\bar{y}2} = \gamma(K) \sum_{i=1}^n \sum_{j>1}^n \bar{y}_{ij} s_i s_j$ $\theta_{\bar{y}2}^* = \frac{n(n-1)}{K(K-1)} \sum_{i=1}^n \sum_{j>1}^n \bar{y}_{ij} s_i s_j$ $\tilde{\theta}_{\bar{y}2} = \frac{n(n-1)}{E(K(K-1))} \sum_{i=1}^n \sum_{j>1}^n \bar{y}_{ij} s_i s_j$
$a_{22} = \sum_{i=1}^n \sum_{j>1}^n E(\bar{x}_{ij}^2)$	$a_{\bar{y}22}$	$a_{\bar{y}22}^* = \frac{n(n-1)}{K(K-1)} \sum_{i=1}^n \sum_{j>1}^n \bar{y}_{ij}^2 s_i s_j$
$a_{23} = \sum_{i=1}^n \sum_{j>1}^n \sum_{\substack{\ell=1 \\ \ell \neq i \\ \ell \neq j}}^n E(\bar{x}_{ij} (\bar{x}_{i\ell} + \bar{x}_{j\ell}))$	$a_{\bar{y}23}$	$a_{\bar{y}23}^* = \frac{n(n-1)(n-2)}{K(K-1)(K-2)} \sum_{i=1}^n \sum_{j>1}^n \sum_{\substack{\ell=1 \\ \ell \neq i \\ \ell \neq j}}^n \bar{y}_{ij} (\bar{y}_{i\ell} + \bar{y}_{j\ell}) s_i s_j s_\ell$
$a_{24} = \sum_{i=1}^n \sum_{j>1}^n \sum_{\substack{\ell=1 \\ \ell \neq i \\ \ell \neq j}}^n \sum_{\substack{m>\ell \\ m \neq i \\ m \neq j}}^n E(\bar{x}_{i\ell} \bar{x}_{\ell m})$	$a_{\bar{y}24}$	$a_{\bar{y}24}^* = \frac{n(n-1)(n-2)(n-3)}{K(K-1)(K-2)(K-3)} \sum_{i=1}^n \sum_{j>1}^n \sum_{\substack{\ell=1 \\ \ell \neq i \\ \ell \neq j}}^n \sum_{\substack{m>\ell \\ m \neq i \\ m \neq j}}^n \bar{y}_{ij} \bar{y}_{\ell m} s_i s_j s_\ell s_m$
$\sigma_2^2 = E(\frac{2}{n(n-1)} \sum_{i=1}^n \sum_{j>1}^n (\bar{x}_{ij} - \frac{2}{n(n-1)} T_2)^2)$	$\sigma_{\bar{y}2}^2$	$\sigma_{\bar{y}2}^{2*} = \frac{2(n+1)(n-2)}{n^2(n-1)^2} a_{\bar{y}22}^* - \frac{4}{n^2(n-1)^2} (a_{\bar{y}23}^* + a_{\bar{y}24}^*)$
$\sigma_{21}^2 = \frac{2}{n(n-1)} \sum_{i=1}^n \sum_{j>1}^n (E(\bar{x}_{ij}) - \frac{2}{n(n-1)} \theta_2)^2$	$\sigma_{\bar{y}21}^2$	- (Zusatzinformationen notwendig)
$\sigma_u^2 = E(\frac{1}{n} \sum_{i=1}^n (U_i - \frac{2}{n} T_2)^2)$	σ_{g1}^2	$\sigma_{g1}^{2*} = \frac{2(n-2)}{n^2} a_{\bar{y}22}^* + \frac{n-4}{n^2} a_{\bar{y}23}^* - \frac{4}{n^2} a_{\bar{y}24}^*$
$\sigma_{u1}^2 = \frac{1}{n} \sum_{i=1}^n (E(U_i) - \frac{2}{n} \theta_2)^2$	σ_{g11}^2	- (Zusatzinformationen notwendig)
$\theta_3 = \alpha \theta_1 + \beta \theta_2; \alpha, \beta \in \mathbb{R}$	$\theta_{\bar{y}3} = \theta_{\bar{y}1} + \theta_{\bar{y}2}$ (S. 146, 174)	$\hat{\theta}_{\bar{y}3} = \hat{\theta}_{\bar{y}1} + \hat{\theta}_{\bar{y}2}$
	$\theta_{\bar{y}3} = \theta_{\bar{y}2} = \theta_{g1} - \theta_{\bar{y}2}$ (S. 176)	$\theta_{\bar{y}2}^* = \frac{n(n-1)}{K(2n-K-1)} \sum_{i=1}^n \sum_{j>1}^n \bar{y}_{ij} (s_i + s_j - s_i s_j)$ bei Stichprobengraphen SVN
	$E(\frac{1}{n} \sum_{i=1}^n \epsilon_i)$ $= \frac{1}{n} \theta_{y1} - \frac{1}{n} \theta_{y2}$ (S. 193)	$\hat{\epsilon} = \frac{1}{K} \sum_{i=1}^n (y_{i1} - \hat{A}_i) s_i$ mit $\hat{A}_i = \frac{n-1}{K-1} \sum_{\substack{j=1 \\ j \neq i}}^n y_{ij} s_j$
$a_{32} = \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n E(x_i \bar{x}_{ij})$	$a_{\bar{y}32}$	$a_{\bar{y}32}^* = \frac{n(n-1)}{K(K-1)} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \bar{y}_{ij} s_i s_j$
$a_{33} = \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \sum_{\substack{\ell>j \\ \ell \neq i \\ \ell \neq j}}^n E(x_i \bar{x}_{j\ell})$	$a_{\bar{y}33}$	$a_{\bar{y}33}^* = \frac{n(n-1)(n-2)}{K(K-1)(K-2)} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \sum_{\substack{\ell>j \\ \ell \neq i \\ \ell \neq j}}^n \bar{y}_{ij} \bar{y}_{j\ell} s_i s_j s_\ell$
$\sigma_{3xu} = E(\frac{1}{n} \sum_{i=1}^n (x_i - \frac{1}{n} T_1) (U_i - \frac{2}{n} T_2))$	σ_{3yg} bzw. $\sigma_{3\bar{y}g}$; speziell mit $x_i = g_i - c_i$ und $U_i = g_i$ (S. 194) oder: σ_{g1}^2 mit $x_i = U_i = G_i$ (S. 177)	$\sigma_{3\bar{y}g}^* = \frac{n-2}{n^2} a_{32}^* - \frac{1}{n^2} a_{33}^*$ (Schätzmöglichkeiten für verschiedene Stichprobengraphen sind schon vorher im Anhang angegeben)

(*) Falls nichts anderes angegeben ist, beziehen sich die Schätzfunktionen auf den Stichprobengraphen SI. Sie sind allerdings auch für die Stichprobengraphen SV, SN und SVN definiert. Andere Schätzmöglichkeiten für den Stichprobengraphen SVN werden auf den Seiten 180, 181 diskutiert.

L E B E N S L A U F

Name: Peter K r e m s e r

Geboren: 6.5.1948 in Opava (CSSR)

Eltern: Alois Kremser, Baumeister
Margarete Kremser, geb. Glombitza, Hausfrau

Familienstand: geschieden

Schulen: 1954 - 1963 Volksschulen in Clarholz, Kreis
Gütersloh, und Bielefeld
1963 - 1968 Friedrich von Bodelschwingh-
Schule in Bielefeld, Aufbau-
gymnasium
22.5.1968 Reifeprüfung

Studium: WS 1968 - WS 1976 Mathematik, Nebenfach
Soziologie, an der Ruhr-Univer-
sität Bochum
2.11.1976 Diplomprüfung Mathematik

Berufliche Tätigkeit: 1.5.1977 - 15.11.1977 Wissenschaftliche Hilfs-
kraft an der Abteilung Mathematik
der Ruhr-Universität Bochum
ab 15.11.1977 Verwalter einer wissenschaft-
lichen Assistentenstelle am
Institut für Statistik und Wissen-
schaftstheorie der Ludwig-Maximi-
lians-Universität München

München, den 23.7.1982

